The Role of Disease-Linked Genetic Variation in the Regulation of Gene Transcription

Soderquest, Katrina

Awarding institution:
King's College London

The Role of Disease-Linked Genetic Variation in the Regulation of Gene Transcription

Katrina Louise Soderquest

Department of Experimental Immunobiology

King's College London

A thesis submitted to the School of Medicine at King's College London for the degree of

Doctor of Philosophy

January 2013

Declaration

I, Katrina Soderquest, declare that this thesis is my own work and that any work performed by others has been acknowledged in the text and properly referenced.

Abstract

Genome-Wide Association Studies have found genetic variation from across the genome to be associated with various diseases and other traits. In many cases, the exact mechanism by which a genetic variant increases or decreases the risk of a particular condition is poorly understood. Many efforts are now underway to combine data on disease-associated genetic variation with other genome-wide data to understand the way in which genetic variation can alter genomic regulation and affect disease risk.

In this project we examine whether disease-associated genetic variation, in the form of single nucleotide polymorphisms, can be found in binding sites for the transcription factors T-bet in Th1 cells and GATA3 in Th1 or Th2 cells. We hypothesise that, in some cases, variation within binding sites for these transcription factors could alter transcription factor binding affinity and subsequent gene regulation. As 'master regulators' of T helper cell lineage commitment, T-bet and GATA3 play a central role in the fate of a naïve T helper cell and the development of an immune response. We find several single nucleotide polymorphisms in our transcription factor binding sites, some of which are near other genomic features such as potential enhancer elements. Furthermore, we find an enrichment of immune related SNPs in T-bet and GATA3 binding sites compared to the total catalogue of Genome-Wide Association Study hits. We then develop a medium throughput assay which combines oligonucleotide pulldown and flow cytometry to test whether such SNPs alter transcription factor binding *in vitro*.

We find three SNPs, rs1465321, rs11135484 and rs1006353 which alter binding of T-bet *in vitro*. Of these, rs1465321 is associated with Crohn's disease, coeliac disease and ulcerative colitis and is in an intron for IL18R1. Therefore, we examine the role of T-bet in IL18R1 expression, IL-18 signalling and two mouse models of disease.

To Rory, for his love and support throughout my entire PhD.

To Jan and Dave, for their love and support throughout my entire life.

Acknowledgements

I would like to acknowledge and thank all of the people listed below for their help and support throughout my PhD:

My primary supervisor for this project was Graham Lord. My other supervisors were Vincent Plagnol, who has provided help and advice on the bioinformatics and statistical genetics side of this work, and Richard Trembath. The other members of my thesis progress committee were Mike Weale, who provided further advice on statistical analysis and Claudia Kemper.

Initial ideas and set-up for OligoFlow experiments were developed with help from Esperanza Perucha. Blood for producing human Th1 cells was collected by and under project ethics of James Canavan. Arrays used to generate figure 5.6 were generated by Refik Gökmen and Kerem Atalar. Binding site data for T-bet and FoxP3 in mouse Th1 cells, as used in figure 4.13, was generated by Refik Gökmen. Adoptive transfer experiment was performed with much assistance from Refik Gökmen. Coeliac experiments were performed jointly with Nick Powell and with assistance from Jon Lo. Slides for Coeliac histopathology were prepared and stained by Chris Evagora. Photographs of slides were taken and scoring and analysis of slides performed by Marjorie Walker.

Thank you also to all the following for their helpful and timely advice: Aditi Kanhere and Richard Jenner for advice on bioinformatic and functional genomic analysis, Rory Blevins for help on R programming, Ian Jackson for all his advice on finding reagents, setting up experiments and finding my way around the laboratory, Arnulf Hertwerk for advice on cloning, Emilie Stolarczyk for help with qPCR and Alfonso Martin-Fontecha for his guidance and clarity of thought at the beginning of my time at Guy's. There is not a single member of the LORD laboratory past or present who has not, at

some point, provided advice, support or a sympathetic ear when the PhD seemed just a little too tough. I am grateful to each and every member and am privileged to have worked with such amazing people. Likewise, I would also like to thank all my friends and family for their support over the past four years.

Beyond the laboratory, I would like to acknowledge support from the BRC funded flow core and thank Susanne Heck, PJ Chana and Helen Graves for sorting cells for my experiments and for their patience. My PhD was funded by the NIHR Biomedical Research Centre at Guy's and St Thomas'.

Much of the computational work within this thesis was made possible by an array of tools made freely available by those that developed them to those that need them. Although I have referenced specific tools in the thesis, I would like to take this opportunity to thank and acknowledge all the people who have contributed to such tools and the open ethos under which they are shared.

Contents

List of Figures

List of Tables

List of Abbreviations

3C Chromosome Conformation Capture

APC Antigen Presenting Cell

bp base pairs

BSA Bovine Serum Albumin

CFA Complete Freund's Adjuvant

ChIA-PET Chromatin Interaction Analysis followed by Paired End Tag Sequencing

ChIP-Chip Chromatin Immunoprecipitation followed by microarray chip

ChIP-Seq Chromatin Immunoprecipitation followed by high-throughput sequencing

CNS Conserved Noncoding Sequence

DC Dendritic Cell

DTT Dithiothreitol

EAE Experimental Autoimmune Encephalomyelitis

EMBOSS European Molecular Biology Open Software Suite

ENCODE Encyclopaedia of DNA Elements

eQTL Expression Quantitative Trait Locus

FACS Fluorescence Activated Cell Sorting

FAIRE Formaldehyde Assisted Isolation of Regulatory Elements

FIMO Find Individual Motif Occurrences

GSI Gamma Secretase Inhibitor

GWAS Genome Wide Association Study

HLA Human Leukocyte Antigen

HRP Horse Radish Peroxidase

IBD Inflammatory Bowel Disease

IFA Incomplete Freund's Adjuvant

IFN-γ Interferon Gamma

ILC Innate Lymphoid Cell

kbp kilobase pairs

LCR Locus Control Region

LD Linkage Disequilibrium

lncRNA long noncoding RNA

LP CI Lamina Propria Chronic Inflammation

LPS Lipopolysaccharide

MACS Model-based Analysis of ChIP-Seq

MEME Multiple Em for Motif Elicitation

MFI Median Fluorescence Intensity

MHC Major Histocompatibility Complex

miRNA MicroRNA

NCBI National Centre for Biotechnology Information

NHGRI National Human Genome Research Institute

NK Natural Killer (cell)

NKT Natural Killer T (cell)

PAMP Pathogen Associated Molecular Pattern

PBMC Peripheral Blood Mononuclear Cells

PBS Phosphate Buffered Saline

PCR Polymerase Chain Reaction

PMA Phorbol 12-myristate 13-acetate

PWM Position Weight Motif

RAG Recombination Activation Gene

SDS Sodium Dodecyl Sulfate

SISSRs Site Identification from Short Sequence Reads

SNP Single Nucleotide Polymorphism

T1D Type 1 Diabetes

TBS-T Tris Buffered Saline with Tween

TCR T Cell Receptor

Th T-helper

Th1Gata3 GATA3 binding sites in Th1 cells

Th2Gata3 GATA3 binding sites in Th2 cells

TLR Toll Like Receptor

TRAP Transcription Factor Affinity Prediction

TSS Transcriptional Start Site

UCSC University of California, Santa Cruz

UTR Untranslated Region

WT Wild-type

1

Introduction

In February 2001 the first drafts of the human genome were published in two separate papers by two separate groups. The International Human Genome Project, comprising researchers from 20 institutions from across six countries officially published their first draft in the journal Nature[1] although a working draft of their work had been released through the UCSC genome server the previous year amid much public attention. At the same time Celera Genomics, a private company, published their first draft in the journal Science.[2] The publicly funded paper estimated only 1.5% of the human genome was protein coding. This figure was subject to caveats but current estimates are similar[3] and the overriding message has not changed in the intervening years. A high percentage of the human genome does not code directly for protein. However, since the publication of the human genome, functional or potentially functional elements have been found in over 80% of the total genome. Researchers have discovered new types of non coding RNA and new elements of gene regulation, culminating most recently in the publications of the Encyclopaedia of DNA Elements (ENCODE) project.[3] The functional diversity of much of the genome highlights the need to understand its regulation at multiple levels. Acquiring the necessary data to fully understand genomic regulation will be a research intensive task. Although helped by the recent development of standardised methods for data generation, sharing and analysis current research also needs a way of prioritising the workload ahead. Which aspects of genomic regulation, in which cellular and physiological contexts, should be studied such that the broad scale work of ENCODE can be applied to understanding, and possibly manipulating, specific mechanisms involved in health and disease? We

can partly prioritise research based on observing and understanding patterns in the genomic data itself. However, we can also prioritise such work based on specific questions arising from other research areas such as immunology where a finely balanced system of T helper cell action and interaction is underpinned by genomic regulation. In this project, we combine functional genomic analysis with an understanding of T helper cell lineage commitment. We do this to examine how genetic variation can usefully be annotated and tested for a potential role in the immune system response and disease mechanism of multiple diseases and other traits.

1.1 Annotating the Genome

1.1.1 Annotating the Genome at the Molecular Level

1.1.1.1 Epigenetics

The exact definition of an epigenetic modification varies. In general an epigenetic modification is a chromatin modification that can be inherited by a daughter cell but does not alter the DNA sequence. However, there is evidence that not all the modifications classed as epigenetic are inherited and even those that can be inherited are reversible.[4] By altering chromatin conformation, epigenetic changes can regulate transcription. The two major types of epigenetic modification are DNA methylation and histone modification. DNA methylation is the addition of a methyl group to a cytosine residue which acts to repress transcription. DNA methylation can prevent the binding of certain transcription factors and also can recruit various repressive complexes. In humans, DNA methylation occurs predominantly at the cytosine residues of CpG dinucleotides. Methylation has been observed at non CpG dinucleotides, particularly at CpA in pluripotent cells, but at far lower levels.[5] Although DNA methylation does not directly alter the DNA sequence, spontaneous deamination of methylated cytosine produces a thiamine molecule which is then passed on as TpG to daughter generations. As a consequence, the occurrence of CpG in most regions of the genome is far lower than would be expected if DNA sequence was random. Regions with higher levels of CpG tend to result from selective pressure against prolonged methylation. Empirically, these CpG islands are

found near genes. High levels of CpG can be associated with both transcription factor binding, which often activates gene expression, and nucleosome occupancy which represses it. The antagonism between these two mechanisms of regulation allows cell specific modulation of transcription.[6] DNA methylation is catalysed predominantly by Dnmt1. Dnmt1 recognises hemimethylated DNA produced during DNA replication and uses this as a template to methylate the newly synthesised strand thus ensuring inheritance of the epigenetic mark. In addition, *de novo* methylation can be catalysed by Dnmt3a and Dnmt3b. Methyl marks can be lost either by active demethylation or through replication in circumstances of reduced Dnmt1 which prevents reproduction of the mark.

The core histone octamer comprises two of each of four different subunits, H2A, H2B, H3 and H4. Each of the four subunits can be covalently modified in various ways including methylation, acetylation, phosphorylation, SUMOylation and ubiquitination. Methylation and acetylation of the H3 subunit are two of the best studied modifications. Of these two modifications, methylation is considered a slower more stable process, marking regions that are either generally open to transcription or generally repressed in the particular lineage of cell under investigation. By contrast, acetylation is a faster process marking regions of active transcription at the time of cell fixation.[7] While different methylation marks can be associated with active or repressed regions, acetylation marks tend only to be associated with active regions. Histone marks have been profiled in a variety of cells, including T helper ($CD4^+$) cells in mouse[8] and human.[9,10] Histone methylation is the addition of one, two or three methyl groups to one of various lysine positions on a histone subunit. Such modifications are denoted by the letter 'K' for lysine followed by the position of the modified lysine residue in the amino acid chain and then me1, me2 or me3 for mono- di- or tri-methylation respectively. The position of modification and the number or methyl units added determines the effect of the modification. For example H3K4me3 is an activating mark whereas H3K27me3 is repressive. Histone methylation marks are added by methyltransferases which contain the catalytic SET domain and their removal is catalysed by the JmjC domain of demethylases. Neither set of proteins is specific for a particular DNA sequence but they are often found in multi-protein complexes and can be guided to specific areas by other transcription factors.[11] A general summary of the best characterised marks and their role is given in table 1.1. It is becoming increasingly apparent that the level of transcription results from an integration of multiple some-

times opposing histone marks. A well known example is the superposition of H3K4me3 and H3K27me3 which poises a gene for transcription.[12] Therefore, the need to integrate multiple datasets is apparent and efforts to do this are ongoing.[13] Acetylation also occurs at lysine residues and the same residues detailed in table 1.1 have the potential to be acetylated to give H3K4ac, H3K9ac and H3K36ac. Only one acetyl group is ever added and the modification is associated with active transcription regardless of which residue is modified.

Modification	Genomic Location	Repressive or Activating
H3K4me1	Enhancers and approximately 1000 base pairs (bp) either side of TSS	Activating
H3K4me2	Promoter regions	Activating
H3K4me3	Promoter regions and Transcriptional Start Site (TSS)	Activating
H3K9me2	10kbp around TSS	Repressive
H3K9me3	10kbp around TSS	Repressive
H3K27me3	Promoters	Repressive
H3K36me3	Downstream of TSS	Activating

Table 1.1: Summary of major histone methylation states.

The remaining histone modifications, phosphorylation, SUMOylation and ubiquitination are also all thought to have a role in transcriptional regulation of gene expression.[14] Phosphorylation is a spatially small modification but has been shown necessary for specific recruitment of other proteins to the chromatin.[15] By contrast, ubiquitin and SUMO are large modifications and may act to help disrupt chromatin in elongation[16] or antagonise other modifications and repress transcription.[17] However, less is known about these modifications and work is still ongoing to understand the mechanisms by which these modifications exert a functional effect.

Nucleosome positioning is also often considered an epigenetic mark. While wound round the histone octamer, DNA is inaccessible and cannot be transcribed. Transcriptional activation, therefore, requires loosening of the DNA. Traditionally, this has been measured by hypersensitivity to DNase treatment as the opened DNA is now accessible to the DNase enzyme. More recently, other techniques have been used to determine DNA loosening such as Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE) in which DNA is cross-linked to nucleosomes and then subjected to sonication. DNA which is not wound round and cannot be cross-linked to a nucleosome is enriched in the

aqueous fraction of a subsequent phenol-chloroform extraction giving results similar to those of DNase hypersensitivity assays.[18]

1.1.1.2 Transcription Factor Binding

A brief search for the term 'regulation of gene transcription, DNA dependant' (GO0006355) in the Gene Ontology yields a list of 4377 gene products. (GO ontology accessed 18th Sept 2012.) Understanding specific examples of and key concepts behind genomic regulation by transcription factors was accelerated by the introduction of Chromatin Immunoprecipitation (ChIP) followed by hybridisation to microarray (ChIP-Chip).[19] ChIP-Chip allowed the analysis of transcription factor binding at promoter sites, and other sites thought important enough to include on a microarray, across the genome. The later introduction of Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) allowed identification of transcription factor binding across the entire genome in a way that was independent of microarray probe choice and thus location relative to coding DNA.[20] It should be noted that there are potential issues with the results obtained from from both ChIP-Chip and ChIP-Seq. Firstly, the results will only be as good as the antibody used in the immunoprecipitation step. The binding affinity and specificity of antibodies for the different transcription factors varies. Because a monoclonal antibody only recognises one epitope it will, in general, yield fewer peaks. A polyclonal antibody, by contrast, will bind multiple epitopes on the transcription factor under investigation and so produce more peaks. Without verifying all peaks using an immunoprecipitation independent method, or by comparing different antibodies for the same transcription factor, it is difficult to say how many of the extra peaks are real and how many are false positives. ChIP-Seq experiments produce a mixture of DNA fragments enriched for those bound to the transcription factor under investigation which must be sequenced and then aligned back to a reference genome. Downstream analysis is complicated by the need to determine the level of sequence enrichment required to call a transcription factor binding peak over background genomic DNA in the sample. Various commercial and non commercial peak calling algorithms have been designed[21] with different strengths and weaknesses[22] and methods for analysing data without calling peaks have also been developed.[3] Further interpretation, downstream of peak calling, is complicated by the complex interplay between transcription

factors and other genomic elements in the regulation of gene expression. Interpretation is also complicated because transient expression of a transcription factor can leave lasting effects through epigenetic modification.[23] Because transcription factors recognise and bind to DNA, binding peaks are often examined for a consensus sequence or Position Weight Motif (PWM). Motifs should be treated with some care as many transcription factors have more than one motif.[24] Furthermore, transcription is regulated by multiple factors binding to the same locus. Therefore, the binding sites of any particular transcription factor may contain sequences recognised by a partner rather than the transcription factor itself. In some cases, a transcription factor may be present at a locus but not bound to the DNA at all and thus not acting through recognition of DNA sequence.[6] For example TCF712 can be tethered to the genome by GATA3.[25] Transcription factors can be relatively promiscuous in the sequences that they will bind. In addition, many more instances of a given motif will occur in the genome than will be bound by the transcription factor. These issues complicate the interpretation of any given DNA sequence with respect to a motif. *In vivo*, transcription factor binding and functional effect will depend heavily on DNA sequence as captured by a PWM but also on binding of other transcription factors and on epigenetic changes and chromatin structure. Nevertheless, sequence alone can predict transcription factor binding and can even do so in a tissue specific way,[26] especially at regions closer to the transcriptional start site (TSS) of a gene.[27]. Presumably, regions close to the TSS rely less on the long range interactions employed by enhancers and other more distal elements. To be maximally predictive, a match to a PWM must be interpreted in the context of other genomic features and transcription factor co-binding. The availability of large databases of PWMs such as JASPAR and TRANSFAC allow the examination of DNA regions for potential transcription factor binding sites. For examining transcription factor co-binding these databases are a useful starting point. The context dependent nature of tissue specific transcription factor binding demonstrates the importance of obtaining genomic information in the appropriate cell type and cell context for the question under investigation. The recent efforts of the ENCODE project have yielded vast quantities of high quality data from which many general genomic lessons can be learnt. An appropriate next step is to apply those lessons to the capture of genomic data in the appropriate kinetic and cellular context to answer specific mechanistic questions about disease basis.

There are many published cases of specific genetic variation altering transcription factor binding at promoter regions. Hummelshoj *et al* demonstrate that a Single Nucleotide Polymorphism (SNP) in the promoter of *SPP1* (also known as *ETA1* or Osteopontin) alters binding of the SP1 transcription factor and this alters transcriptional activity as examined by luciferase assay.[28] SP1 binding at the *IL10* promoter is also altered by a SNP.[29] Less is known about the functional consequences of altered transcription factor binding at the genome-wide level. Interpreting data at this level is complicated by the non-linear relationship between transcription factor binding and gene expression in many cases.[30] However, by profiling binding of the general transcription factor Pol II and the specific transcription factor NF-κB, Kasowski *et al* were able to show substantial differences in transcription factor binding strength between ten different lymphoblastoid cell lines. Some of these differences correlated with both the presence of a SNP and altered mRNA expression.[31] More recently, the ENCODE project has found that the combination of chromatin state and transcription factor binding at promoters can be quantitatively correlated with RNA production. The ENCODE project also found that disease associated SNPs are enriched in non-coding functional elements including transcription factor binding sites.[3] Furthermore, a recent Genome-Wide Association Study (GWAS) found a SNP associated with bipolar disorder within a binding site for the Oestrogen receptor ER-α[32] but they did not directly show altered binding.

So far, much of the work to examine whether a SNP alters transcription factor binding and quantity of transcript produced has been based on correlations between ChIP-Seq or ChIP-Chip data, SNP data and variation in mRNA levels with genotype. However, there are more direct ways of testing for the functional effects of some SNPs. In the case where the SNPs in question are exonic, then allele specific expression can be assayed in heterozygotes for that SNP. In this case, the SNP appears in the mRNA transcript and so the relative levels of mRNA transcribed from each allele of the SNP can be compared. Levels of mRNA are normalised to the signal for genomic DNA to control for potential differences in probe used, as the alleles will be in a 1:1 ratio in genomic DNA. To apply a similar approach to the whole genome, cDNA can be hybridised to a SNP array and relative abundance of the mRNA of two different alleles can be examined at any region where the donor is heterozygous.[33] SNP arrays can also be used to assay the relative abundance of two alleles in samples that have undergone ChIP.[34] However, although these techniques are potentially

very powerful, they are somewhat limited by the fact that SNP arrays do not contain every single SNP. Furthermore, allele specific techniques require a large amount of material from any one donor. As such, these techniques tend to use either cell lines or immortalised cells, which raises some issues about the applicability of the results to primary cells. These issues are discussed later.

1.1.1.3 Chromatin Interactions

Evidence for chromosomal looping between enhancer and promoter regions was first published in 1989.[35] As techniques for unbiased genome-wide analysis of three dimensional chromosomal conformation have developed, the importance of regulation at this level has received increasing attention. Techniques for mapping the three dimensional conformation of chromatin have evolved from the Chromosome Conformation Capture (3C) technique.[36] This technique involves cross-linking proximal DNA regions, cutting the DNA by restriction enzyme, ligating together the ends of the cross-linked products, reversing the cross-links and then determining the sequence of the ligated products by Polymerase Chain Reaction (PCR).[37] The cross-link step does not discriminate between DNA sections that are functionally interacting and those that happen to be in close proximity at the time of fixation. Therefore, truly interacting DNA sections must be registered as an enrichment against a background of randomly ligated products. The 3C technique examines interactions between specific loci but variations on the 3C technique, including 4C and 5C, have been developed. While 3C examines the interactions between specific loci, 4C examines all possible interactions between a specific locus and the rest of the genome. In this case, ligated products are circularised once the cross-links have been reversed. Primers are designed against two sections of the locus of interest allowing any interacting sections of the genome to be amplified by inverse PCR. These amplified sections can then be analysed by large scale sequencing or microarray.[38] In 5C, a library of ligated fragments is generated in the same way as for the 3C technique. However, primers are then designed to multiple different loci of interest. Tags of extra sequence are added to each primer. These tags allow further amplification of the product of any primer pair by universal primers. A further round of amplification is then used to prepare a library suitable for large scale sequencing or microarray analysis.[39]

It is also possible to incorporate ChIP into the 3C type methodologies to ex-

amine the role of a transcription factor in the DNA interactions under investigation. By combining high throughput sequencing, these methodologies have also been adapted for unbiased analysis of chromatin interactions across the whole genome with the advent of Hi-C[40] and Chromatin Interaction Analysis followed by Paired End Tag Sequencing (ChIA-Pet).[41]

Data on three dimensional chromatin structure have been conceptually useful in understanding the annotation of the genome. For example, epigenetic marks may imply a region of DNA is an enhancer element but without three dimensional information we cannot be sure which genes it is an enhancer for. Typically, it is assumed that enhancers act on their nearest gene and this is often the case but not always. For example, the regulatory elements of the alpha-globin locus act on the gene *NME4* which is 300 kilobase pairs (kbp) away, without affecting genes between the two regions. This regulation is through physical interaction between enhancer regions at the alpha-globin site and the promoter of *NME4*.[42] Because the *NME4* promoter and the promoter regions of the alpha-globulin genes (of which there are two copies in a healthy individual) compete, the authors found that deletion of one or both of the alpha globin genes increased expression of NME4. This finding highlights the complex role that genetic variation can have on gene expression. Recent results from the ENCODE project suggest the promoter regions of the majority of genes are involved in tissue specific 'multi-gene' interactions.[3] Chromosomal looping occurs both at the *IFNG* locus and the locus containing the genes for IL-5, RAD50, IL-13 and IL-4 (hereafter referred to as the Th2 cytokine locus). Furthermore, interactions between the *IFNG* locus and the Th2 cytokine locus, which are on different chromosomes, have also been reported in naïve $CD4^+$ cells.[43]

1.1.1.4 Conservation Across Species

One major way of annotating the genome has been to find Conserved Noncoding Sequences (CNS) which are sequences that are conserved between species. This is based on the theory that elements that are functionally useful are under negative selection and so are more likely to be retained through evolution. Comparisons between the mouse and human genome for example have been used in annotating multiple loci across the genome including the *IFNG* and Th2 cytokine loci. Although this relatively simple technique has been very useful, we must be somewhat cautious. Comparison is often made, for example, be-

tween mouse and human. Wide scale analyses of these genomes suggest that the sequence and usage of some elements, such as promoters, are highly conserved between these species. However, the sequence or the way in which other elements are used, such as enhancers and binding sites for the transcriptional repressor CCCTC-binding factor are more variable between species.[44] In addition, regulatory function can be conserved across species without necessarily conserving sequence.[45] Furthermore, in the case of the immune system, continual changes driven by host-pathogen interaction may render this system less likely to be conserved compared to other systems.

1.1.2 Annotating the Genome at the Functional Level

1.1.2.1 Transcription Level

To add another level of annotation to the genome, we can assess the impact of genetic variation at a region on a downstream product. Much work has now been done to correlate genetic variation with levels of transcription through analyses of expression Quantitative Trait Loci (eQTL.) Although computationally intensive and subject to issues around multiple testing, genome-wide comparisons between SNPs and gene expression can yield many useful insights. However, as with genomic analyses, accurate transcriptomic analysis requires the appropriate tissue or cell type.[46,47]

It is also important to realise that although RNA levels result from the integration of multiple genomic signals, they themselves will be moderated before translation. Such modification may include the action of other RNAs and post-transcriptional processing. We cannot assume that mRNA levels will always correlate with downstream protein levels and other outputs. A recent study suggested that approximately 40% of variation in protein levels was explained by variation in mRNA levels, with much of the remaining variation explained by the rate of translation.[48,49] In terms of eQTL analysis, variation has been found that affects the 5′ Untranslated Region (UTR) of a transcript and thus affects protein expression at the translational level.[50] There are practical trade-offs to be made in transcriptomic analysis. We can analyse the transcriptome in a very specific cell type that we think may be affected by a DNA variant, such as effector T cells, to try and obtain a high signal to noise ratio in that particular

cell type. However, this approach might leave us with low cell numbers. Alternatively, we can use a broader category of cell type such as whole blood or Peripheral Blood Mononuclear Cells (PBMC). However, transcriptional changes in one small subpopulation of a more heterogeneous population of cells may get lost in the noise of our data. In the first case, cell numbers are often increased by immortalisation of primary cells to form a cell line. Many transcriptomic analyses have been performed on the Epstein-Barr immortalised cells produced by the HapMap and 1000 Genome Project. The immortalisation process can itself affect genome methylation and gene expression.[51]. Systematic studies suggest transformation affects over half the genes in the genome but that the effects are relatively small in magnitude.[52] Information on transcriptional variation relevant to primary cells can be obtained from such lines[53] even though experimental artefacts may produce false positives. However, it is still important to be aware of the limitations of such cell lines. For the analysis of heterogeneous populations of cells, work is ongoing to deconvolute signals from microarrays bioinformatically.[54]

1.1.2.2 Transcription Product

Non coding functional classes of RNA, such as rRNA and tRNA, have long been known. However, RNA molecules are involved in a wider range of processes and other classes such as microRNAs (miRNAs), long noncoding RNAs (lncRNAs), piwi-interacting RNAs and enhancer RNAs have been described. The functions and mechanisms of many of these types of RNA are still being elucidated. One of the better defined classes are miRNAs. Mature miRNAs are RNA molecules of approximately 21 nucleotides that act to reduce expression of their targets by targeting mRNA for degradation or translational inhibition. Some miRNAs are known to have an important role in gene regulation. In a non-immune context, the first discovered miRNA family member let-7 is conserved across multiple species including mouse and human and is important in development.[55] In an immune context miR-146a targets STAT1. Loss of this miRNA results in loss of immune tolerance and early death in a mouse model as a consequence of impaired Treg function.[56] Dysregulation of miRNA function has been implicated in disease. Loss of function mutations in the miR-96 locus causes nonsyndromic autosomal dominant progressive hearing loss.[57] Some cases of Feingold syndrome are caused by heterozygous deletion of the miR-

17-92 cluster.[58] Feingold syndrome is an autosomal dominant condition which results in an unusually small head, abnormalities of the fingers and toes, learning disabilities and often heart or kidney defects. MicroRNA dysregulation is also implicated in cancer and various cardiac conditions.[59,60] Genetic variation can impact on miRNA action by altering the expression level or sequence of the miRNA itself, by altering the seed sequence of an mRNA target or by creating a seed sequence in a novel target. Although no mendelian conditions have been found to result from alteration of a target mRNA sequence, complex diseases have been associated with both types of variation.[60–62]

Long noncoding RNAs are typically defined as non coding RNA molecules that are over 200bp long. Genome-wide transcript analyses have identified thousands of lncRNAs which are now catalogued online.[63] The importance of this functional class in the regulation of immune genes was highlighted recently. Tmevpg1, a lncRNA that is approximately 170kbp upstream of the *IFNG* locus has been reported to have a role in Th1 cells.[64]

1.1.3 Annotating the Genome at the Clinical Level

1.1.3.1 The Genome-Wide Association Study

By understanding and annotating the genome at the molecular and functional level, we can begin to understand how genomic variation can impact on host phenotype. However, the analysis can work both ways: by studying host phenotype, we can understand more about the genome at the molecular level. Such studies often involve analysis of genetic variation associated with specific phenotypes such as a particular disease. To date, one of the most widely studied types of genetic variation between individuals is the Single Nucleotide Polymorphism (SNP). There is a vast literature documenting the role of various SNPs in health and disease mainly resulting from two types of investigative approach candidate and genome-wide.[65] In candidate-based approaches, researchers pick a small number of SNPs for study, based on prior biological knowledge, and work from these to find disease mechanisms. More recently developed genome-wide approaches, such as Genome-Wide Association Studies (GWAS), analyse thousands of SNPs across the entire genome in a non-directed fashion. Such genome-wide techniques do not require prior hypotheses or par-

ticular biological knowledge and so they can highlight previously unconsidered biological pathways in a hypothesis-generating manner.

The most common form of GWAS is the case-control GWAS, in which the allelic frequencies of SNPs from a population with a particular illness or other trait are compared to SNPs from a control population. Although this method can be applied to any trait and GWAS have been performed for non disease traits such as mathematical ability,[66] most GWAS analyse a diseased versus a healthy population. Defining the relevant populations can be problematic in practice. As GWAS itself has highlighted, diseases can be heterogeneous at the molecular level, with multiple pathologies resulting in the same clinical outcome.[67] Understanding disease at this molecular level is one way in which many researchers hope to start a move towards personalised medicine.[68] GWAS results can also be skewed by variability in non-disease related traits such as ethnicity and gender.[65,69] However, GWAS have now been conducted on a wide range of diseases yielding huge amounts of data and some important insights. The first GWAS, conducted in 2005, highlighted the role of complement in age-related macular degeneration.[70] Other successes include the role of the autophagy pathway in Crohn's disease.[71]

In many cases of Mendelian conditions, linking genetic variation back to molecular mechanism at the genome level is relatively straight forward. For Mendelian conditions, the disease will be caused by the presence of a highly penetrant SNP or other genetic variant which can be linked to the disease by analysing inheritance of the trait within families. Segregation analysis can be used to narrow down candidates for the causal genetic variant of the disease to the causal variant itself and any variants genetically linked or fixed to it. Such candidates can then be tested for functional effect. Functionally, the genetic variation underlying Mendelian conditions alters transcriptional product. Often, this is because the variation is in a protein coding sequence and results in dysfunctional protein, although Mendelian conditions can occur as a result of mutations in other contexts. Mutations at splice sites will also alter the protein product of a gene and, as already discussed, mutations in other transcriptional products such as miRNAs can also cause Mendelian conditions.

For complex diseases linking genetic variation, molecular mechanism and disease state is more challenging. The concept of causality in complex disease is more complicated than in cases of Mendelian conditions. Complex diseases do

not have one cause; rather they result from an interplay between multiple genetic and environmental factors. In the case of complex disease, a causal SNP will not independently cause the disease but exert a far more subtle effect. For example a SNP may increase or decrease risk of disease or moderate disease response to a particular set of environmental parameters. Because of this incomplete penetrance we cannot readily find genetic variants linked to a complex disease in the same way as for a Mendelian condition. GWAS for complex conditions have yielded a wealth of data on associations between different SNPs and diseases. However, there are relatively few cases in which any particular moderation of disease risk or course has been linked to just a few testable SNPs, or where one causal SNP has been identified. More commonly, a GWAS will highlight a region of highly co-inherited SNPs that are more common in a diseased population than a healthy control population and are therefore described as associated with the disease. SNPs can be co-inherited and the r^2 value gives a measure of how much the frequency at which two SNPs are co-inherited deviates from the frequency of co-inheritance we would expect if those SNPs were inherited independently of each other.[72] This is a measure of their linkage disequilibrium (LD). A high r^2 value shows that the SNPs deviate strongly from a pattern of independent co-inheritance: they are in strong LD and likely to be co-inherited. Practically, this is useful as we can infer information about some SNPs from others and do not need to analyse every SNP in a genome to find associations. Such work has been further aided by the efforts to categorise LD patterns across the genome in multiple individuals from multiple ethnicities in the HapMap project.[73,74] However, it also means that a SNP associated with a particular condition is not necessarily causal at the molecular level: it may simply be in high LD with a causal SNP. There are also issues with relying on data from HapMap to give every SNP in a population. The third iteration of HapMap, HapMap 3, genotyped SNPs in 1184 individuals across 11 populations.[74] In the example of the CEU population, which are individuals with Northern and Western European ancestry living in Utah, genotyping was performed on 55 trios to make a total of 165 samples. If we assume that the 110 parents have completely different genomes then the project is trying to capture the variation in the population across 220 samples for each autosomal chromosome. Thus, any SNP at a minor allele frequency of 5% in the population (the cut-off at which a SNP is considered a common SNP) has an approximately 0.0013% chance of being absent from the samples. This would suggest approxi-

mately 20 of the 1.6 million SNPs tested might have been missed if all the SNPs were on the threshold of being common variants. However, a greater proportion of rare variants, with a minor allele frequency of less than 5% would have been missed. Furthermore, the HapMap data relies on genotyping of specific SNPs rather than using full sequence data. Although the number of SNPs genotyped is extensive, there is still potential for SNPs to be missed without direct sequencing of the genome.

Finding the causal SNPs, linked to complex disease, may be possible genetically through the use of fine mapping and sequencing over large cohorts of patients and controls. Such fine mapping studies, such as those using the ImmunoChip, have found that some genomic regions probably contain more than one causal variant for certain diseases. For example, analysis of coeliac disease, using the ImmunoChip, found multiple independent signals in the loci surrounding, among others, *IRF4*, *STAT4*, *RGS1* and *TAGAP*.[75] Such a fine level of detail would have been missed in early GWAS where, for reasons of cost, studies genotyped fewer SNPs per sample. However, mechanistically the discovery of multiple independent signals in some loci is perhaps not surprising given the extensive interactions between cis and trans factors in regulating gene expression.

Finding the functional consequences of disease-associated SNPs can be challenging. Many disease-associated SNPs have been found to reside in so called 'gene deserts', regions devoid of any coding sequence. Given our current understanding of genome regulation, as discussed above, it is likely that most of these 'deserts' do have function. Understanding that function in the appropriate cell type, at the appropriate time and in the appropriate context is a huge task but progress is being made. For example, a gene desert on chromosome nine was recently found to contain multiple enhancer regions. SNPs associated with coronary artery disease in one of these enhancer regions were shown to alter STAT1 binding and subsequent expression of CDKN2A and CDKN2B following stimulation by IFN-γ.[76] Disease-associated SNPs are also found at the 5′ and 3′ ends of genes, in introns and at splice sites. Both synonymous and nonsynonymous coding region SNPs are also found. Although the presence of a nonsynonymous SNP immediately suggests a causal mechanism, not all result in dysfunctional protein.[77] An immediate focus on nonsynonymous SNPs to the neglect of other SNPs can miss SNPs with important regulatory effects.[78]

Efforts to predict nonsynonymous SNPs that negatively affect protein function are ongoing.[79,80] Interestingly, nonsynonymous SNPs seem to have a similar size of effect, in terms of disease association, to synonymous SNPs.[81]

Genome-wide analysis of disease-associated SNPs in functional elements has been explored recently by the ENCODE project. This research found an enrichment of SNPs from the National Human Genome Research Institute (NHGRI) GWAS catalogue in transcription factor binding sites versus all SNPs from the 1000 Genome Project (12% NHGRI versus 6% 1000 genome).[3] Most of the ENCODE work was performed on cells lines and to better understand any potential biological mechanism underlying this enrichment, it is important to study individual cases in the appropriate cell types to the phenotype in question.

It is important to note that GWAS have only explained a small proportion of the heritability of the traits that have been studied. There is lively debate as to the source of this 'missing heritability' centring around three possibilities. There may be small contributions from many common SNPs that each have effect sizes below the limit of detection by GWAS. Alternatively, larger contributions from moderately penetrant SNPs, in which each SNP is too rare in the population to be found by a large scale screen across the population, may have a role. Finally, the missing heritability may result from an under-appreciation of the effect of gene-environment interactions. There are many arguments for and against these ideas[82] and it is likely that different diseases will have different contributions from each of these effects. Arguments for the idea of rare variants include work from large scale sequencing such as the 1000 Genome Project that suggest any given genome will contain a high number of rare or private variants. Current estimates are that any genome will differ from the reference genome at approximately 10,000 nonsynonymous SNPs and a further 10,000 synonymous SNPs.[83] Under the rare variant model, we would expect some GWAS hits to tag, not one causal SNP, but several different rare variants in different people. However these synthetic associations do not explain all the success that GWAS have had. As some researchers have pointed out, the quest to find missing heritability results, in part, from confusing heritability with genetic contribution.[84] Heritability is often measured from family studies. For example one measure, the relative risk between siblings, measures the likelihood that an individual will develop a particular disease, given that their sibling is affected, compared to the average risk of any member of the

population developing that disease.[85] This provides an estimate of heritability but to take it as a measure of genetic contribution would be to ignore similarities in environment between the two siblings.[86] Indeed, it is worth noting that, since the development of GWAS, other 'Wide Association Studies' have been performed such as the 'Environmental-Wide Association Study'[87] and 'the Epigenome-wide Association Study'.[88] These studies highlight the need to generate and integrate data across multiple features. The epistatic effects which result from gene-environment interactions and also from gene-gene interactions complicate the analysis of complex conditions. At the genetic level, functional interaction between two SNPs has been observed in psoriasis.[89] At the gene-environment level, functional interaction between genetic disposition and an obesogenic environment has been implicated in the rise of obesity and type II diabetes (TIID).[90,91]. As a more specific example, SNPs have been found that that associate with increased incidence of oesophageal squamous cell carcinoma, but only in alcohol drinkers.[92]. Thus, a genetic variation may 'cause' a condition but only under certain environmental conditions. The difficulties with measuring and quantifying environmental factors and calculating gene-environment interaction may explain some of the 'missing heritability' problem.

1.1.4 Public Data and Open Access

The story of the human genome sequencing effort is often pitched as a race between two groups, The International Human Genome Project consortium, a publicly funded group and the private company Celera Genomics. Concerns that private success and the patents that would follow would limit access to and use of data on the genome, coupled with rapid improvements in sequencing techniques, pushed the public consortia to publish earlier than originally planned.[93] As a result, the two groups published their data together in 2001.[1,2] By this point, the public consortium had already made a working draft of the genome publicly available. A section of the Nature paper published by the consortium outlines its commitment to collaboration and timely unrestricted access to data. This has set the tone for much of the genomic research that has followed. However, the sheer volume of data now available presents its own challenges in terms of data storage, data analysis and data visualisation.

Many of the commonly used databases and tools for genomics are held and

maintained either by the American National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) or by the European Bioinformatics Institute and UK's Sanger Centre under the Ensembl project (http://www.ensembl.org/index.html). There are discrepancies between the data held by the two groups, for example, the list of Ensembl genes differs slightly from the list of RefSeq (NCBI) genes.[94] However, there is also huge overlap in the data provided by each database and better characterised genes will have both an Ensembl identification starting ENSG (ENST for the gene transcript) and a RefSeq annotation starting NG (NM for the transcript). A Consensus Coding Sequence identification is assigned when there is general consensus across the major groups that a DNA sequence is a protein coding gene. There are numerous other sources of genomic information but most will not be discussed here. Of importance to the questions addressed in this project dbSNP, another NCBI database, provides information on SNPs including published clinical relevance. Because repositories such as dbSNP obtain their information from data submitted by researchers, in the spirit of open access, the quality of work performed to find and analyse any particular SNP varies. As such, dbSNP also provides information on the source of information for each SNP and on the confidence in the accuracy of the annotation provided. Another major source of information on other genomic features is the recently released ENCODE data.[3]

Both Ensembl and NCBI provide tools for analysing genomic data. In addition, the UCSC genome browser was introduced in the International Genome Project consortium publication[1] and remains an important tool in visually aligning multiple datasets.[95] Tools for more advanced analysis, many of which can be readily used by researchers unfamiliar with bioinformatic tools, have been developed more recently. While the many modules of the Bioconductor toolkit[96] require some basic knowledge of the R programming environment,[97] other tools such as Galaxy[98–100] work through a user-friendly web based interface. A review of some of the available tools has been performed by Nielson *et al*.[101]

1.1.5 Systematic Functional Annotation - The ENCODE Project

The ENCODE project aimed to identify all functional elements of the human genome.[102] It started with a pilot phase in September 2003 which analysed 44 discrete regions from across the genome. These regions were chosen to repre-

sent a range of functional characteristics. In total, the 44 regions covered approximately 1% of the human genome. To assay such elements in a systematic way, that could be shared and reproduced across laboratories, the ENCODE consortium used existing standards and, where appropriate, developed new standards for data acquisition, analysis and sharing. In some cases, the consortium also used multiple experimental protocols for obtaining data on the same biological function. For example, transcript analysis was performed by hybridisation of RNA to tiling arrays, sequencing of 5′ ends of transcripts and analysis of Expressed Sequence Tags. By the time of its publication in 2007,[103] the pilot phase data had already been released, primarily through the UCSC genome browser. Among its many findings, the pilot phase concluded that much more of the genome is transcribed than is protein-coding. The pilot phase also showed that, in contrast to what was previously thought, there are many examples of non-coding functional genomic elements that are not constrained across mammalian evolution.

The full ENCODE project, which investigated the full genome, was published in September 2012 as 30 papers across the journals Nature, Genome Research, Genome Biology and BMC Genetics. As previously mentioned, ENCODE was able to assign functionality to just over 80% of the human genome.[3] This included elements such as protein coding genes, promoters and enhancers but also regions bound by transcription factors and sites of DNase hypersensitivity. ENCODE studied such features across different cell lines and primary cells. For the purposes of the ENCODE project, cells were grouped into three tiers. Tier one, which consisted of three cell lines, was the most extensively studied. The three cell lines for this tier were K562, a cell line of erythroleukemia lineage, GM12878, a B lymphoblastoid cell line and the H1 hESC stem cell line. These cells were subject to the full range of ENCODE techniques. We must be slightly cautious in the data generated from these cell lines. K562 is a cell line isolated from Chronic Lymphocytic Leukaemia and has an abnormal karyotype.[104] GM12878 has been immortalised with Epstein-Barr virus which may have affected some aspects of its genomic function as previously discussed. Furthermore, these cell lines do not encompass all cell types or tissues. A wider range of cells is found in tier three, including T cells of different $CD4^+$ subsets, but these have not been examined by every ENCODE technique. Some functional elements specific to these cells alone, or to those cells not covered at all by ENCODE, may still be discovered. Thus the 80% figure given by ENCODE

may be subject to alteration in the future.

However, ENCODE has provided a wealth of useful data. In particular, much work has been undertaken to find general signatures for different genomic regions in terms of histone binding, DNase hypersensitivity and transcription factor binding. Such general signatures can then be applied to those cells and tissues for which we have limited data, allowing us to infer function from the data that we do have. For example, Th1 and Th2 cells were studied by ENCODE as a tier three cell line and, therefore, only subjected to DNase hypersensitivity assays. However, Arvey *et al*[105] develop a framework for predicting cell-type specific transcription factor binding based on DNase hypersensitivity and DNA sequence. This framework could be applied to the DNase hypersensitivity data in Th1 and Th2 cells. Natarajan *et al*[106] perform similar work to predict gene expression from DNase hypersensitivity data, combined with transcription factor binding motifs.

Of particular interest to the work in this project, Schaub *et al*[107] use the genomic information found by ENCODE and other sources to functionally annotate SNPs from the NHGRI GWAS catalogue. They also determine which SNP from all the SNPs in high LD with a disease-associated SNP is most likely functionally relevant. Through this method, Schaub *et al* find potential functional effects for up to 80% of reported associations. However, they do not try to confirm the effect of these functional SNPs *in vitro* or *in vivo*. Furthermore, given its reliance on ENCODE data, their work does not necessary use cells and tissues specific to disease. Some of the work by Schaub *et al* is based on a database, RegulomeDB which was produced by another ENCODE team.[108] RegulomeDB can be used to find the functional relevance of any SNP across all the data in the database, which includes all the ENCODE data and data from other sources. The likelihood that any queried SNP has functional relevance is ranked based on whether it forms part of an eQTL, whether it is in a transcription factor binding site or peak of DNase hypersensitivity and whether it alters a know transcription factor binding motif. Although the data included in the database is extensive (962 datasets over more than 100 tissues and cell lines), it is not exhaustive. If data for the binding of a particular transcription factor or the DNase hypersensitivity of a particular cell type does not exist then it will not be in the database and will not be included when ranking queried SNPs. Use of the database carries this limitation, although the authors express

the expectation that the database will grow with time as further whole genome analyses are conducted and added to the database.

1.2 Overview of T Helper Cell Immunology

1.2.1 Activation of the Immune Response

To infect the body, a pathogen must first cross one or more of various barriers such as the skin or mucosal lining of the airways or gut. Once past these initial defences, the pathogen will encounter elements of the innate immune system. Early immune activation is triggered, in large part, by molecules that are generally expressed across a range of pathogens but are either not expressed by cells of the host or are spatially separated from components of the innate immune system. This includes components of the bacterial cell membrane such as Lipopolysaccharide (LPS)[109] and nucleic acid structures used by some viruses such as double stranded RNA.[110] These Pathogen Associated Molecular Patterns (PAMPs) are recognised by pattern recognition receptors on host cells such as the Toll Like Receptors (TLRs)[111] and by components of the complement system.[112] TLRs are located on both the cell surface and in endosomes and lysosomes. They are mainly found on cells of the innate immune system but can also be found on cells of the adaptive immune system and on cells of non-haematopoietic lineage.[113] Signalling downstream of TLR engagement uses the same pathways as the IL-1R family[114] and includes molecules such as MyD88 and members of the IRAK family. Engagement of TLRs in conjunction with antigen triggers the maturation of dendritic cells (DC) which then migrate to the secondary lymphoid tissues. Dendritic cells are the only type of Antigen Presenting Cell (APC) able to present peptide to naïve T cells.[115] Although produced in the thymus, naïve T cells also migrate to the secondary lymphoid tissue where they will, during an immune response, encounter activated DCs. The way in which the adaptive immune response is activated will broadly result from an integration of three signals, antigen recognition, costimulation and local cytokine environment.

1.2.2 The T Helper Cell Lineages

1.2.2.1 Overview

Most mature T cells express the α and β subunits of the T Cell Receptor (TCR) and can be divided into one of two types, cytotoxic T cells which express the surface molecule CD8 ($CD8^+$ cells) and helper T cells which express the surface molecule CD4 ($CD4^+$ cells). Other subsets such as $\gamma\delta$ T cells exist but will not be detailed here. Naïve $CD4^+$ cells are activated through TCR and costimulation. Following rapid proliferation, they produce cytokines to direct the immune response, activate and promote isotype switching in B cells[116–118] and activate macrophages.[119] $CD4^+$ cells are generally categorised into one of several subsets according to their cytokine expression pattern and transcription factor expression profile. These subsets are Th1, Th2, Th17, regulatory T cells (Treg) and follicular helper T cells (Tfh). In general terms, a viral infection will initiate a Th1 response and Th1 cells will then produce, among others, the cytokine IFN-γ.[120] There is some debate as to the primary role of the Th2 response but it is thought to be protective against helminths[121] and other extracellular pathogens and it results in IgE production. Various studies have demonstrated that administration of IgE based antibodies could be an effective treatment against solid tumours[122–124] and inverse correlations between certain allergies and certain forms of cancer have been published.[125,126] This suggests that a strong Th2 response may be protective against some cancers. However, a general role for the Th2 response in cancer surveillance and protection remains ill-defined and controversial. Th2 cells produce IL-4, IL-5 and IL-13. Feedback mechanisms have been shown to skew an immune response to either Th1 or Th2, enhancing one response while simultaneously repressing the other. Th17 cells have been characterised more recently and are generally thought to be a response to extracellular bacteria and fungi: they produce, among others, the cytokine IL-17.[127,128] Tregs produce IL-10 and are generally considered anti-inflammatory cells which dampen the immune response elicited by other T helper cell subsets.[129] Tfh cells help B cells: they are needed for the formation and maintenance of germinal centres.[130] Although Tfh cells are often considered a separate subset, the development of Tfh cells intersects with that of the Th1 lineage.[131,132] Autoimmune and inflammatory diseases often arise from an overactive or improperly triggered Th1 or Th17 response. It should be noted that the $CD4^+$ cell subset model

is somewhat simplistic. Other subsets such as Th9 and Th22 cells[133] have been described and there is growing evidence that the cell subsets are not as fixed or clearly defined *in vivo* as they are after long term cell culture. These issues are discussed in more detail later.

Each subset can be defined by a 'master regulator', a transcription factor which directs the expression programme for the subset and can inhibit other subsets. The expression of each master regulator is not necessarily confined to its own subset. Furthermore, a master regulator is traditionally defined as the transcription factor required for the generation of a particular subset but questions have arisen over how we define the subsets in light of recent data. Nevertheless, the notion of a master regulator is conceptually useful. The master regulators for Th1, Th2, Th17, Treg and Tfh are T-bet, GATA3, RORγT, FoxP3 and Bcl-6 respectively.[134]

1.2.2.2 Activation and Skewing of the Adaptive Immune System

To become active, a naïve $CD4^+$ cell must receive two main signals which are provided in the secondary lymphoid tissue by a matured DC. The TCR of the $CD4^+$ cell must bind its cognate antigenic peptide presented by the class II Major Histocompatibility Complex (MHC) on the DC. In addition, costimulation must be provided, usually by interaction between CD28 on the surface of the $CD4^+$ cell and CD80 or CD86 (also known as B7-1 and B7-2) on the surface of the DC. The requirement for costimulation helps ensure that T cell responses are only triggered in an inflammatory setting and not in response to self peptide. Negative selection in the thymus results in autoreactive T cells being deleted or undergoing a process of receptor editing such that they are no longer autoreactive. However, some autoreactive T cells do reach maturity and further mechanisms in the periphery must exist to prevent them causing autoimmune conditions. T cells which engage TCR in the periphery without costimulation become anergic. Although $CD28^{-/-}$ mice can still produce some T cell responses, they are markedly diminished.[135] As a further checkpoint on activation, various inhibitory molecules exist such as CTLA-4. CTLA-4 also binds CD80 and CD86 but with higher affinity than CD28, particularly in the case of CD80.[136] The mechanism by which CTLA-4 inhibits T cell activation is still widely debated.[137] Possible mechanisms include downregulation of CD80 and CD86 on APCs by CTLA-4 expressing Treg cells or reduction of costimulation due to competition

between CD28 and CTLA-4 for the same ligands. Either way, this added checkpoint helps prevent T cell activation by self- peptide. Knockout of CTLA-4 in mice results in fatality within 3-4 weeks.[138] CD80 and CD86 can modulate IL-4 and IFN-γ production.[139]

Antigen signal strength can also affect the type of response: it is proposed that low and high strength signals promote a Th2 type response and middle strength signals promote a Th1 type response.[140] Analyses of antigen signal strength and costimulation suggest that the interplay between the two can influence commitment to the Th1 or Th2 lineage as varying signal strength is measured against a backdrop of thresholds set by costimulation.

A greater factor in $CD4^+$ lineage commitment is the local cytokine environment during activation. Commitment to the Th1 lineage is generally initiated by IFN-γ concurrent with TCR signalling. Although DCs can produce IFN-γ, they are not the major producers of this cytokine. TCR signalling can result in very low levels of IFN-γ secretion from the naïve $CD4^+$ cell itself but extra IFN-γ is provided by Natural Killer (NK) cells, which can also be recruited to the lymph node under conditions of inflammation and infection[141]. In one model of Th1 lineage commitment, a preliminary burst of IFN-γ allows an initial small upregulation of the Th1 master regulator T-bet. T-bet then upregulates the inducible unit of the IL-12 receptor[142] following extinction of the initial TCR signal.[143] Under resting conditions, DCs do not produce IL-12 but they will do so when activated through their TLRs. In fact, TLR engagement tends to produce mature DCs that promote a Th1 rather than a Th2 response[114,115] and thus the most appropriate response to the invading pathogen is initiated. IL-12 activates the transcription factor STAT4[144] and this stabilises T-bet expression and reinforces the Th1 response by further upregulating T-bet, IFN-γ and various other Th1 related genes such as Runx3 and Hlx. Of note, this model implies two distinct waves of T-bet expression modulated by different pathways.[143] However, such linearity has been questioned by recent data suggesting a high level of redundancy between IL-12 and IFN-γ in promoting the expression of T-bet.[145] IFN-γ (in addition to TLR stimulation) can induce IL-12 production from macrophages which is an important feedback mechanism to continue the Th1 response after the activated T cell has left the lymph node and migrated to the site of inflammation.[146]

The Notch pathway is also important in Th1 versus Th2 lineage commit-

ment.[147,148] $RBPJ_k$ (also known as CSL) binds constitutively at the promoter of *TBX21*, the gene for T-bet, and can recruit transcriptional repressors to this region. However, such repressors can be displaced by the intracellular domain of Notch1 which then recruits transcriptional activators. Cells deficient in intracellular Notch1 either due to knockdown by antisense construct or treatment with gamma secretase inhibitor (GSI) are less able to produce IFN-γ. Gamma secretase is required to cleave total Notch1 and release the intracellular domain. Cells treated with GSI have reduced expression of T-bet. However, expression of T-bet and of IFN-γ is restored through expression of a constitutive intracellular domain of Notch1.[149]

Th2 lineage commitment is initiated by IL-4, although the initial source and mechanism is still not fully defined. The lymph node contains very little if any IL-4 but TCR signalling on its own can upregulate IL-4 transcription in an activating $CD4^+$ cell. This small amount of IL-4 can then act in an autocrine and paracrine manner, signalling through STAT6, to upregulate GATA3 expression and reinforce a Th2 programme.[150] Furthermore, signalling through Lck, a Src family kinase which is a key effector in TCR signalling, acts to inhibit Th1 type responses. Kemp *et al*[151] showed that naïve $CD4^+$ cells from a mouse model in which Lck is deleted in $CD4^+$ cells in the periphery were less able to produce IL-4 when cultured in Th2 polarising conditions. This was accompanied by increases in Runx3, IFN-γ and T-bet production. Epigenetically, Th2 cells from these mice were normal at the Th2 cytokine locus but had permissive acetylation marks at the IFN-γ locus which are usually only present in Th1 cells. This suggests that defects in this aspect of TCR signalling result in decreased initial commitment to the Th2 lineage not due to a failure to actively promote the Th2 lineage but due to a failure to actively repress Th1 commitment. This idea is further enforced by the finding that Lck loss did not effect Th1, Th17 or Treg lineage commitment in conditions designed to skew to those responses. Furthermore, naïve cells have been shown to default to Th2 when activated under neutral conditions.[152] This suggests a situation *in vivo* in which Th2 might be the default immune response as has already been proposed.[153] Microbial or fungal antigens which need to be counteracted with a Th1 or Th17 type response induce IL-12 or IL-23 production from DCs leading to a dominant Th1 or Th17 type response. By contrast, an absence of this stimulation, in cases where such responses are not required, will allow a default to Th2.

The Notch pathway has also been shown to have a role in Th2 lineage commitment. In addition to activating T-bet, the intracellular domain of Notch1 can also upregulate GATA3 and this increase in expression is independent of IL-4 and STAT6.[148]

On leaving the lymphoid tissue, the activated T cells migrates to the site of infection through various chemokine receptors. Th1 cells express CXCR3 and CCR5, both of which are T-bet targets. Th1 cells promote B cell antibody switching to IgG2a and activation of macrophages via the 'classical' pathway. Th2 cells express CCR3, CCR4 and CCR8, and promote B cell antibody switching to IgE and IgG4 in human or IgG1 in mouse. They can also promote eosinophil activation and activation of macrophages via the 'alternative' pathway.

1.2.2.3 A Note on the 'Textbook' Lineage Definitions

The classification of cells into distinct subsets began with the observation by Mosmann *et al*[120] that T cell clones could be placed into one of two groups based on their cytokine profile. The Th1 and Th2 subsets were shown to respond to different types of pathogen and modulate the immune response in different ways, as discussed above. Both subsets have their own 'master regulator' transcription factor and once established in the body each subset uses various mechanisms to repress the other lineage. Other types of $CD4^+$ cell which express different cytokine profiles, have since been discovered, requiring the addition of new subsets to the model such as Th17 and Treg. However, questions have been raised as to whether the classification of distinct subsets really reflects the true nature of the $CD4^+$ cells in an immune response *in vivo* and whether the responses are as fixed as first thought. Repeated stimulation and long term culture yields the distinct subsets seen by Mosmann *et al* and different autoimmune conditions can be attributed to the overactivity of one particular subset. Nevertheless, IFN-γ production has been observed from cells skewed to Th2 *in vitro* and *in vivo*.[154,155] Furthermore, although IFN-γ acts to repress early Th17 lineage commitment,[156] it can later be produced by Th17 cells themselves. As described by Lee *et al*, such cells may then lose production of IL-17 in some conditions[157] and these observations raise important questions about lineage definitions. Th2 cells cannot, by definition, produce IFN-γ if we define Th2 on the basis of IL-4 production in the absence of IFN-γ production; cells which are not producing IL-17 are not, by definition Th17 cells if we define

Th17 on the basis of IL-17 production. These issues highlight problems with the subset model. At the level of transcription factors, GATA3 may be the master regulator of the Th2 lineage but it is expressed in Th1 cells.[158] Furthermore, the master regulator of the Th1 lineage, T-bet, can be found in Treg cells.[159] As previously discussed master regulators remain poised for transcription even in opposing lineages suggesting potential for reexpression. Furthermore, Th1 and Tfh cells may share a common developmental stage[131] and conditions of low IL-2 can increase Bcl-6 expression in Th1 cells and promote expression of some Tfh genes.[132]

1.2.3 Key Cytokines in the Classical Th1, Th2 Cell Model

1.2.3.1 IL-12

Sources of IL-12 include DCs, monocytes and macrophages.[160] These cells must be stimulated to produce IL-12 through a mixture of TLR ligation,[114,115] CD40-CD40L interaction with already activated T cells,[161] and exposure to cytokines such as IFN-γ.[162] The main targets of IL-12 are NK and T cells but IL-12 has also been shown to have effects on macrophages[162] and DCs themselves.[163] IL-12 consists of two chains p40 and p35. Of these, it is the availability of p35 that limits the availability of functional IL-12. Upon stimulation p40 is more highly expressed than p35. Spare p40 can bind another subunit, p19, to form IL-23 a cytokine important in Th17 lineage commitment.[164] In symmetry to IL-12, the formation of IL-23 is restricted by the availability of the p19 subunit. The IL-12 receptor consists of two chains: IL12Rβ1 is constitutively expressed on naïve T cells, whereas IL12Rβ2 must be activated by T-bet[165] before a cell can respond to IL-12. Interestingly, the constitutive IL12Rβ1 can also bind to IL23R to form a functional receptor for IL-23 in the same way that the IL-12 subunit p40 can bind to a different subunit p19 to form the IL-23 protein.[166] This highlights the molecular interplay between the Th1 and Th17 lineages in early T cell lineage commitment. IL-12 primarily activates the STAT4 pathway and, through this, upregulates and helps to stabilise IFN-γ production in a Th1 type response. (See IFN-γ section for further details.) IL-12 also works synergistically with IL-18 to produce IFN-γ independently of TCR engagement in pre-activated T cells and in NK cells. IL-12 also activates Lck and the p38 pathway.[167]

1.2.3.2 IFN-γ

A fully functional IFN-γ protein is a homodimer of two IFN-γ peptide chains[168] and is the only known member of the type II interferon family. Main sources of IFN-γ are Th1 cells, cytotoxic $CD8^+$ T cells and NK cells, although B cells, Natural Killer T (NKT) cells and APCs can also secrete this cytokine.[146] As suggested by its original name, Macrophage Activating Factor, IFN-γ can act on and is important in the maturation of macrophages but it is also very important for the Th1 pathway. Naïve T cells require TCR stimulation for initial activation but antigen experienced T cells can produce IFN-γ in both a TCR dependant and TCR independent manner.[169] TCR independent IFN-γ production results from stimulation by IL-12 and IL-18, with the combination of these cytokines working synergistically for maximal IFN-γ expression.[170] IL-12 and IL-18, also induce IFN-γ from non-T cells such as NK cells. IFN-γ production is downregulated by IL-4 and TGF-β.[171] The IFN-γ receptor consists of two IFN-γR1 units and two IFN-γR2 units. Signal transduction is usually limited by availability of the IFN-γR2 unit. However, although low levels of IFN-γR2 are required for signal transduction and proliferation in response to IFN-γ, high levels of IFN-γR2 and the consequentially high levels of signalling are antiproliferative and promote apoptosis.[172] The role of relative signal strength in determining outcome is an important theme in immunology which can allow small variations in signal to exert broad effects.

IFN-γ activates STAT1. $STAT1^{-/-}$ mice are developmentally normal but cells from these mice cannot respond to IFN-γ and the mice show a similarly impaired response to the microbial pathogen *Listeria monocytogenes* as IFN-$\gamma R1^{-/-}$ mice.[173] However, while STAT1 is important in IFN-γ signalling, not all IFN-γ responsive genes are direct STAT1 targets. For example, *IL1R1* is a STAT1 target which can be downregulated in macrophages by IFN-γ. This leaves macrophages unable to upregulate other non-STAT1 targets in response to IL-1 such as various matrix metaloproteinases.[174] In order to activate transcription, the STAT proteins must oligomerise to form dimers or trimers. Although IFN-γ usually exerts its effects through STAT1 homodimers, it can also signal through the STAT1:STAT2:IRF-9 trimer which is more typically associated with type I interferon signalling. IFN-γ signalling results in various phosphorylation events throughout the standard JAK-STAT signalling cascade.[175] However, to be fully effective, the STAT1 homodimer must also be phosphorylated at a

serine residue by a separate signal which can originate from various sources including IL-2, IL-12 or TNF-α.[176]

In terms of broader effects, IFN-γ promotes immune responses by upregulating surface expression of MHC Class I and by promoting changes to the proteasome that increase the diversity and the quantity of peptides presented to $CD8^+$ cells.[146] IFN-γ can also upregulate class II MHC.[177] As previously mentioned, IFN-γ can induce IL-12 production from macrophages. It can also prime macrophages by increasing expression of TLRs and inhibiting negative feedback loops that would otherwise dampen macrophage response to PAMPs. For example, the IL-10 pathway is activated during TLR signalling and acts to dampen the response to TLR signalling. However, IFN-γ can act to antagonise IL-10 signalling in this setting.[178] IFN-γ drives the Th1 response, which is generally considered pro-inflammatory. However, in doing so, it antagonises other T helper cell responses such as the Th17 response. It is notable that IFN-$\gamma^{-/-}$ mice develop worse responses in some models of autoimmunity which were originally thought to be Th1 driven but are now believed to be predominantly Th17 driven. These models include Experimental Autoimmune Encephalomyelitis (EAE). The interplay between IFN-γ and T-bet is also important in immune modulation as T-bet represses the Th17 programme[179] and is expressed in some Treg cells to allow homing to sites of inflammation[159].

Given its importance in immunology, the IFN-γ locus has been well studied. The complexity in its regulation provides an example of how multiple pathways and mechanisms can be integrated to fine-tune an output such as gene expression. As shown by Soutto *et al*, a small region around the *IFNG* transcriptional start site (TSS) (-565bp to +64bp) can promote IFN-γ production in a lineage specific way in $CD4^+$ cells but not in $CD8^+$ cells.[180] The proximal *IFNG* promoter region contains multiple monomeric Brachyury sites that are T-bet responsive and show cooperation between sites.[181] A full Brachyury site contains the palindromic sequence CACxxxxGTG and is named after the Brachyury protein, the founder member of the T-box family to which T-bet belongs. A monomeric or half Brachyury site consists of CAC of GTG only. In $CD4^+$ cells, T-bet is the primary driver of IFN-γ production. However, T-bet is partly functionally redundant with another transcription factor Eomesodermin (Eomes) in $CD8^+$ cells which may explain the promoter region discrepancies between $CD4^+$ and $CD8^+$ cells.[182] Enhancer regions are found in introns one and three of

the *IFNG* locus but these are not able to confer lineage specificity. Indeed a much larger region around the *IFNG* locus, stretching approximately 90kbp either side of the protein coding region, is required for robust IFN-γ expression that is also lineage specific in both $CD4^+$ and $CD8^+$ cells.[183] Regulatory elements in this region have been found through comparing conservation of sequence between species. Critical enhancers at CNS regions 22kbp (CNS-22) and 34kbp (CNS-34) bind T-bet[184] and CNS-34 is involved in DNA looping at the locus. Such looping also involves other elements within the *IFNG* locus as well as an element in an intron of *IL16*, the nearest gene to *IFNG* in humans but a gene that is not conserved in mice.[185] The looping also requires the CCCTC-binding factor and cohesin.[186] Use of these elements is dynamic, changing throughout the course of T helper cell lineage commitment. DNase profiling shows a changing chromatin structure as a naïve $CD4^+$ cell develops towards either the Th1 lineage, where chromatin must be accessible and IFN-γ must be highly expressed or the Th2 lineage, where chromatin must be tight and IFN-γ must be highly repressed. Furthermore, the range of regulatory mechanisms employed allows the expression of IFN-γ to reflect the integration of TCR dependant and independent pathways.[170] In addition to interactions between regulatory elements around the *IFNG* locus, interactions are also seen across chromosomes between the *IL4* and *IFNG* loci.[43] Further details on the epigenetic control of the locus, such as the other cis elements involved, have been reviewed by Balasubramani *et al*.[165]

Important transcription factors that can bind at the *IFNG* locus include STAT4, T-bet and NF-κB. The NF-κB family member RelA is activated downstream of IL-18. RelA co-binds at the *IFNG* locus with STAT4 in TCR independent IFN-γ induction integrating the signals of IL-12 and IL-18. RelA is also important in TCR dependant IFN-γ production.[170] $STAT4^{-/-}$ mice can produce IFN-γ but IFN-γ production by $STAT4^{-/-}$ cells skewed to Th1 conditions *in vitro* is severely impaired.[187] IFN-γ production is also severely impaired in $CD4^+$ from $T\text{-}bet^{-/-}$ mice, although IFN-γ production by $CD8^+$ is less impaired[188] most likely due to the redundancy between T-bet and Eomes in these cells.[182]

1.2.3.3 IL-18

IL-18 was originally identified as Interferon Gamma Inducing Factor.[189] It functions synergistically with IL-12 to induce TCR independent IFN-γ production

in Th1 cells.[170,190] IL-18 is a more potent inducer of IFN-γ than IL-12 but, unlike IL-12, IL-18 cannot, on its own, induce a Th1 type response from a naïve cell on primary stimulation. Once a naïve cell has been stimulated through the TCR and becomes an effector memory Th1 cell then IL-18 and IL-12 can promote IFN-γ production independently of further TCR engagement on this cell.[190] IL-18 is produced as an inactive precursor, pro-IL-18, by a wide range of cells including macrophages, B cells, T cells, DCs, keratinocytes and intestinal epithelial cells. Pro-IL-18 is then usually cleaved to form mature IL-18 by caspase-1 before being secreted from the cell.[191] However, uncleaved pro-IL-18 can be found in human blood and is secreted by human PBMCs.[192] It is then possibly cleaved extracellularly by chymase. The chymase enzyme has been shown to cleave IL-18 although at a different site to produce a mature IL-18 fragment with less activity than IL-18 that has been cleaved by caspase-1.[193] The IL-18 receptor can be expressed on a wide variety of cells including Th1 cells, NK cells, macrophages, B cells, neutrophils and some endothelial and epithelial cells.[194] The IL-12 and IL-18 signalling pathways crossregulate: IL-18 signalling upregulates the IL12Rβ2 subunit but downregulates the IL-18 receptor itself whereas IL-12 signalling upregulates the IL-18 receptor.[195] The receptor comprises IL18R1 also called IL18Rα and IL18RAP (Receptor Accessory Protein) also called IL18Rβ. The IL18R1 subunit can weakly bind IL-18 and has limited if any signalling capacity. The necessity for a further subunit, originally called AcPL but later renamed to IL18RAP, was first demonstrated in 1998 by Born *et al*.[196] This finding was confirmed by mouse knockout model in 2005 by Cheung *et al*.[197] IL18RAP cannot bind IL-18 at all but, when bound to IL18R1 the resulting complex can both bind to and allow signalling through IL-18. IL-18 signalling activates the MyD88/IRAK/TRAF6 pathway which ultimately results in NF-κB activation and translocation.[190,198] IL-18 can also activate the MAPK cascade in mature T cells.[199] IL-18 signalling can be inhibited by a decoy protein, IL-18 binding protein, that binds IL-18 and prevents it from binding to the IL-18 receptor. Signalling through IL-18 is therefore, very similar to signalling through IL-1β and IL-33. A summary of some of the IL-18, IFN-γ and IL-12 signalling pathways discussed can be seen in figure 1.1.

In addition to IFN-γ induction from Th1 cells, IL-18 is also needed for optimal IFN-γ induction from NK cells. As with Th1 cells, IL-18 can synergise with IL-12 to optimally induce IFN-γ from these cells. Furthermore, a role for IL-18 in priming NK cells such that they can upregulate IFN-γ in response to IL-12 has

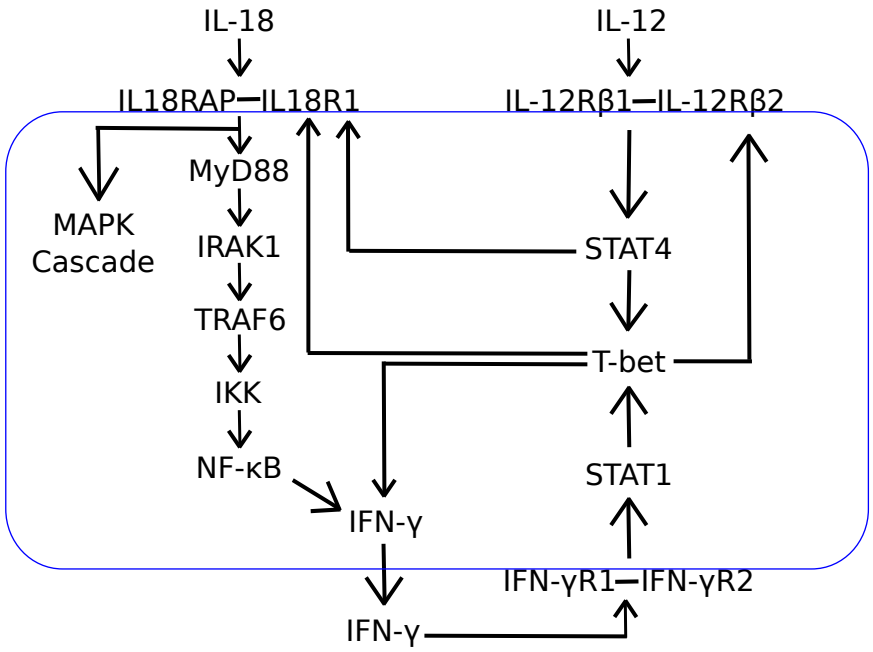

Figure 1.1: Main components in IL-12, IL-18 and IFN-γ signalling pathways. - Schematic of some of the signalling interactions between IL-12, IL-18 and IFN-γ in Th1 and other cell types. Many components omitted for clarity.

been described.[200] Natural killer cells from IL-18 deficient mice are impaired in IFN-γ production compared to wild-type cells when stimulated with IL-12 *in vitro*. This defect is not seen when IL-18 signalling is blocked at the time as IL-12 administration to wild-type cells suggesting a role for IL-18 signalling upstream of IL-12 in this setting, a reversal of the kinetics seen in Th1 cells. IL-18 signalling also enhances NK cell cytotoxicity.[201] Both IL18R1$^{-/-}$[202] and IL18RAP $^{-/-}$[197] mice fail to increase cell killing in response to IL-18 compared to wild-type controls. These separate studies demonstrate that both IL18R1 and IL18RAP are required to mediate this effect.

Initial studies of the IL-18 receptor demonstrated its expression on and importance for Th1 cells and its absence on Th2 cells.[195] However, the role of IL-18 signalling in other settings involving CD4$^+$ cells has been published. It has been suggested that low levels of IL18R1 expression on naïve cells can allow general inflammatory signalling that, in the absence of IL-12, can skew a cell towards the Th2 type lineage. Yoshimoto *et al*[203] demonstrated that activation of CD4$^+$ cells in the presence of IL-18 and IL-2 resulted in a higher percentage of IL-4 producing cells and a lower percentage of IFN-γ producing cells than cells incubated with IL-2 alone. However, this result most likely reflects the plasticity

of naïve cells during activation and the ease with which they can be artificially encouraged to produce an array of different cytokines at this stage. IL-18 does not have a specific role in IL-4 production. Epigenetically, the *IL18R1* locus follows a similar trend to the *IFNG* locus. The chromatin at the *IL18R1* locus is accessible in naïve $CD4^+$ cells as shown by hypersensitivity to DNase treatment. The locus is then activated in Th1 cells as seen by the addition of H3K4me2, H3K4me3 and acetylation histone marks whereas the chromatin is tightened to become actively repressed in the Th2 state.[204] The poised state of the locus in naïve cells could allow some IL18R1 expression. IL18R1 expression has been demonstrated on unstimulated cells and it has been suggested that such expression is seen prior to cell activation but is then downregulated as a result of CD3 ligation.[190,204–206] In the presence of IL-12, IL18R1 is then upregulated to be expressed at high levels, while it is further downregulated by IL-4.[204,206] Such results must be interpreted with some caution as most experiments have been performed on cultures enriched for all $CD4^+$ cells including naïve, effector and memory cells. Few experiments have been performed on the pure populations of naïve cells that can be obtained by sorting or using an antigen specific system such as the D011.10 or OT-II systems on a background where all other T cells have been deleted through deletion of one of the Recombination Activation Genes (RAG). However, the studies raise some interesting questions. Are IL18R1 and IL18RAP definitely expressed on a pure population of naïve cells and at what levels compared to Th1 and Th2? If they are expressed, are they expressed at high enough levels to allow signalling? Furthermore, if IL-18 signalling can promote other lineages in the very early stages of $CD4^+$ cell lineage commitment then what subset of conditions, in terms of relative timings and strengths of IL-12 and IL-18 signals, allow for non-Th1 pathways to be pursued and how relevant is this to disease?

Fully polarised Th1 cells do not produce IL-4 in response to IL-18. However, there are some studies to suggest that fully polarised Th1 cells can be induced to produce the typically Th2 associated cytokine IL-13, by the presence of IL-18 in the absence of IL-12, on secondary stimulation in humans[207] and mice.[208] The levels of IL-13 produced are lower and later than in typical Th2 cells but are nonetheless clearly present and are thought to have a role in asthma. It is also thought that IL-18 can induce IL-17 production under certain conditions. Studies by Lalor *et al*[209] show that $CD4^+$ mouse cells that are stimulated with αCD3, αCD28 and IL-18 produce some IL-17. IL-17 production is synergisti-

cally increased by the addition of IL-23. Furthermore, the combination of IL-18 and IL-23 can induce IL-17 production on unstimulated cells whereas this can not be achieved by either cytokine alone. These results add weight to the argument that IL-18 acts as a general inflammatory signal that can be directed in its outcome by other cytokines. However, these experiments were performed on a bulk $CD4^+$ population. Therefore, we cannot determine the necessary order of IL-18 and IL-23 presence and TCR engagement. We cannot distinguish IL-17 production from naïve cells undergoing primary stimulation and effector cells undergoing a further round of stimulation. Furthermore, although the authors use a model of EAE and show increased IL18R1 in diseased mice, they do not directly correlate IL18R1 expression with IL-17 production in $CD4^+$ cells. Therefore, the role of IL18R1 expression on $CD4^+$ cells in Th17 cell development and maintenance remains to be fully defined. However, it has been shown that IL18R1 expression on APCs is needed for the generation of Th17 in a mouse model of EAE.[210] This same study also suggests a role for IL18R1 that is independent of IL-18 as discussed below.

In the general model, IL18R1 and IL18RAP bind to form a complex through which IL-18 signals. However, there are some reports showing that the IL-18 receptor has functions which are independent of IL-18. For example, $IL18R1^{-/-}$ mice are resistant to EAE induction whereas $IL\text{-}18^{-/-}$ mice are susceptible. However this is not thought to result from deficiency of IL18R1 on the $CD4^+$ cells but on APCs. In the model, IL18R1 deficient APCs were less able to secret the p40 subunit of IL-23.[210] This does suggest that another ligand, possibly of the IL-1 superfamily, can signal through IL18R1. Interestingly, in humans IL-37, also named IL-1H or IL-1F7, has been shown to bind to IL18R1.[211] However, the EAE studies were performed in mice and this protein has not been identified in mice. Furthermore, IL-37 is generally seen as an anti-inflammatory molecule which antagonises IL-18 signalling.

It is unclear as to whether IL18R1 and IL18RAP are co-regulated in terms of their expression. The two loci neighbour each other on chromosome two in humans and chromosome one in mice which suggests potential for co-regulation and evidence for this has been found by some studies.[204] However, in mice, some reports suggest that while low levels of IL18R1 are expressed on naïve T cells, IL18RAP is completely absent until IL-12 stimulation.[212] This raises issues about the capacity of IL-18 to induce non Th1 associated function, such as the

production of IL-4, from naïve $CD4^+$ cells in the absence of IL-12. Furthermore, the two genes seem to vary in their relative requirement for various transcription factors and chromatin modifications for expression.[23,213] As with much of the work in the field, most of the controversies are most likely due to differences in species, strain, timepoint of assay and definition of naïve cell used. Greater clarity over the kinetics of IL-18 signalling with respect to these parameters would be useful for future research. Of note, in one of the coeliac GWAS, Hunt *et al* found expression of IL18RAP but not IL18R1 in whole blood correlated with genotype at one of their significantly associated SNPs, rs917997.[214] In this case, the risk allele correlated with lower expression of IL18RAP. At a genetic level, this strongly implicates IL18RAP and IL-18 signalling in coeliac disease. At a mechanistic level it, again, raises questions around the extent to which IL18R1 and IL18RAP are co-regulated and whether co-regulation varies across different cell subsets. For example, we might ask whether expression of IL18R1 would vary with genotype if we specifically studied a Th1 population rather than whole blood.

Optimal expression of the IL-18 receptor requires STAT4 and T-bet.[213] STAT4 binds the proximal promoter of the *IL18R1* locus.[23] T-bet binds the proximal promoter of *IL18RAP* but not *IL18R1*.[158] Interestingly, however, Thieu *et al* found reduced but detectable expression of IL18RAP in cells from T-bet deficient mice whereas they found no detectable IL18R1 in the same cells. This argues for a central role of T-bet in IL18R1 expression and, by extension, demonstrates the importance of transcription factor binding at locations other than the proximal promoter. It also suggests that IL18R1 and IL18RAP are, at best, only partially co-regulated. By contrast, Balasubramani *et al* found that IL18R1 expression was impaired but not abrogated in T-bet deficient mice.[170] Therefore, the role of T-bet at this locus merits further investigation. Interestingly, the nearest upstream gene to IL18R1 is IL1RL1 which encodes the protein T1/ST2/IL-33R. IL-33 is thought, in many ways to be the Th2 equivalent of IL-18.[194]

1.2.3.4 IL-4, IL-5 and IL-13

Unlike the IL-12 receptor, a functional IL-4 receptor is expressed on naïve $CD4^+$ cells.[150] On T cells, IL-4 receptor is a heterodimer of IL4Rα and the common gamma chain, so called because it is a subunit not just of the IL-4 receptor but also of the receptors for IL-2, IL-7, IL-9, IL-15 and IL-21. Small amounts of

multiple cytokines including IL-4 but also IFN-γ can be produced by cells on activation through the TCR. However, a move from low level promiscuous cytokine production to the sustained IL-4 production of a differentiated Th2 cell requires the Th2 master regulator GATA3.[215] GATA3 does not bind at the *IL4* promoter and once a cell is fully committed to the Th2 lineage, IL-4 can be expressed without GATA3, although GATA3 is still required for maximal IL-4 production.[216] GATA3 does bind to a distal enhancer element of the *IL4* locus along with NFAT,[217] a transcription factor that is expressed in both Th1 and Th2 cells but with different targets in the two lineages. It is possible that IL-4 production requires GATA3 dependant chromatin remodelling and subsequent NFAT binding rather than GATA3 binding itself which would explain the variation in GATA3 requirement with time. NFAT also binds the *IL4* promoter as does the Th2 lineage restricted transcription factor c-Maf.

GATA3 is required to maintain expression of both IL-13 and IL-5 in committed Th2 cells.[216] The genes for IL-4, IL-5 and IL-13 are all found close together in the Th2 cytokine locus on chromosome five in humans and chromosome 11 in mice. The genes for IL-4 and IL-13 neighbour each other and are separated from the gene for IL-5 by the gene for Rad50. Rad50 is not a Th2 associated protein but is ubiquitously expressed and involved in DNA repair. Although GATA3 does not bind the promoter of *IL4*, it does bind throughout the region including at the *IL5*[218] and *IL13*[219] promoters and the first intron of the *IL4* gene.[220] In naïve cells, the *IL4* locus is transcriptionally poised with permissive and repressive histone marks and heavy DNA methylation. The locus becomes less methylated as cells move towards Th2 lineage commitment and another early role of GATA3 may be to antagonise DNA binding protein MBD2. MBD2 binds methylated DNA but such binding is prevented by GATA3.[221]

Various other epigenetic changes are seen as cells progress from naïve to Th2 and these have been reviewed by Ansel *et al*.[150] Of note, many of the cis regulatory elements which control expression of IL-4 and IL-13 stretch across two distinct regions, one containing the *IL4* and *IL13* genes themselves and a Locus Control Region (LCR). An LCR is a region which co-ordinately enhances expression of more than one gene in a tissue specific fashion. In this case, the LCR is a 25kbp region in the 3' end of the *RAD50* gene[222] (fig. 1.2). Assays using the 3C technique have shown looping at the locus which brings *IL-4*, *IL-13* and *IL-5* together. This looping is seen even in non lymphocyte cells such as

fibroblasts but in naïve T cells the LCR is also in the loop poising the locus for transcription. In Th2 cells the interactions become stronger. Both GATA3 and STAT6 have been shown to have a role in the chromatin loop formation and maintenance.[223]

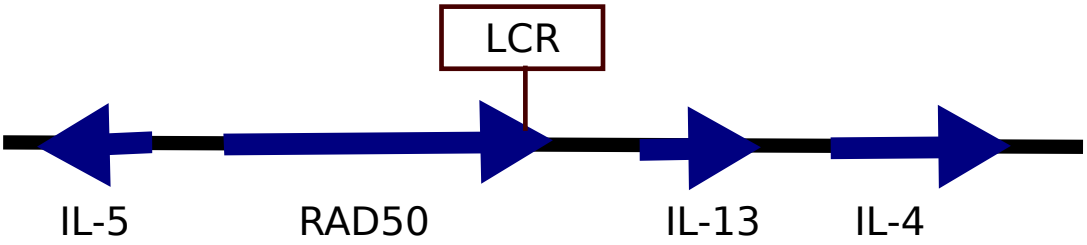

Figure 1.2: Relative arrangement of genes and LCR at Th2 cytokine locus. - Schematic of chromosomal region containing genes for IL-4, IL-5, IL-13 and RAD50 showing relative location of LCR.

In addition to its role in Th2 cells, IL-4 is produced by mast cells, basophils, eosinophils and NKT cells. IL-4 also promotes antibody switching, increases the expression of class II MHC in B cells and modifies the expression of adhesion molecules on endothelial surfaces.[224]

IL-13 signals through a dimer of IL13Rα1 and IL4Rα which is expressed on a number of cells including epithelial cells, macrophages and eosinophils but not T cells.[225] IL-4 can also signal through this receptor and, for this reason, the receptor complex is often called the type II IL-4 receptor (the type I receptor is the heterodimer of IL-4Rα and the gamma common chain described above). IL-13 shares many functions with IL-4 which, given the common receptor, is not surprising. However, distinct functions of IL-13 have been characterised. IL-$4^{-/-}$ mice can expel the gastrointestinal parasite *Nippostrongylus brasiliensis* with similar efficiency to wild-type mice but IL-$13^{-/-}$ mice have impaired goblet cell function and take far longer to expel these parasites.[226] Furthermore, IL-13 is a more potent inducer of fibrosis than IL-4. This is seen in chronic infection with *Schistosoma mansoni* which promotes a continual Th2 response. Although the Th2 response is mounted to fight the infection, constant exposure to Th2 cytokines leads to liver scaring. Fibrosis is reduced in IL-4 deficient mice compared to wild-type controls but is reduced substantially further in wild-type mice treated with an IL-13 inhibitor.[227]

IL-5 signals through the IL-5 receptor, a dimer of IL5α and IL5β. While IL5α is only used for IL-5 signalling, IL5β is shared with receptors for IL-3 and GM-

CSF. IL5Rα is expressed by and is important for certain subset of B cells such as murine B1 cells. The IL-5 receptor is also expressed on eosinophils and IL-5 is required for parasite induced eosinophilia. However, IL-$5^{-/-}$ mice have baseline numbers of eosinophils that are comparable to wild-type controls.[228]

1.2.4 Master Regulators

1.2.4.1 T-bet

Named for its original discovery as a **T-B**ox motif protein **E**xpressed in **T** cells, T-bet was first discovered in 2000 as a protein that bound the *IL2* promoter and was expressed in Th1 but not Th2 cells.[229] A full crystal structure of T-bet has not yet been published but the DNA binding sections of human TBX1,[230] TBX3[231] and TBX5[232] and of the T-box family founder member *Xenopus* Brachyury[233] have all be crystallised. These structures suggest that T-box proteins bind to DNA as dimers across a region of approximately 25bp, with contacts made in both the major and minor grooves. The relative importance, in maintaining a DNA binding dimer, of the protein-protein contacts at the interface of each monomer versus the protein-DNA contacts that hold each monomer onto the DNA seems to vary between the factors.[231] The binding of the T-box family members as dimers in addition to various *in vitro* analysis would suggest a preference for palindromic consensus motifs based around CACxxxxGTG for DNA binding. This motif is often referred to as the Brachyury consensus motif. However, half sites (Brachyury half-sites) are common in T-box regulated promoters *in vivo* implying that a half site is sufficient. The crystallisation of the TBX5 T-box on a natural half-site promoter suggests that T-box proteins still bind as dimers at half-sites and that the dimerisation is not an artefact of crystallisation on a synthetic palindromic sequence. However, the T-box proteins do not dimerise in free form- they are monomers when in solution.[232] A consensus motif for T-bet has only just been published[234] and work is till ongoing to determine the specifics of T-bet binding to DNA.

T-bet is encoded by the gene *TBX21*. As a family, the T-box proteins tend to be involved in development and mutations in several of the T-box family members are strongly associated with disease. Mutations in *TBX1* are associated with DiGeorge syndrome, mutations in *TBX3* cause Ulna-mammary syndrome and

mutations in *TBX5* cause Holt-Oram syndrome. There are no known mendelian conditions resulting from mutations in the *TBX21* gene itself. Associations between SNPs in *TBX21* and asthma[235] and diabetes[236] in Japanese patients have been reported but not replicated. To date, disease-associated polymorphisms in *TBX21* have not been found from any large scale GWAS that meets the criteria for inclusion in the NHGRI GWAS catalogue, although disease-associated polymorphisms have been found in *TBX2*, *TBX3*, *TBX4*, *TBX5*, *TBX15* and *TBX20* (http://www.genome.gov/gwastudies/index.cfm).

T-bet is best known as a transactivator of the *IFNG* locus. However, as its title as 'master regulator' suggests, it has far wider functions including IFN-γ independent suppression of the Th2 lineage.[229] T-bet also binds to and influences expression of a number of other genes involved in Th1 function including *CXCR3*[237] and *IL18RAP*.[158] T-bet is thought to bind at many genes but only influence expression at a subset of its targets.[238] This was originally shown by Beima *et al*, who also demonstrated that T-bet targets were similar across B, T and NK cells. However, this study used cell lines rather than primary cells for the initial ChIP experiments so such data should be interpreted with some caution. More recently Kanhere *et al* have shown that T-bet binds some genes both proximally and distally and that the expression of these genes is more likely to be regulated than those genes bound only proximally by T-bet.[234]

T-bet knockout mice are viable and mostly normal, with normal T and B cell development although they have fewer peripheral NK and NKT cells and impaired NK and NKT cell development.[239] Such mice are more susceptible to infection with *Mycobacterium tuberculosis* than wild-type mice[240] and cannot clear infection by *Leishmania major*[188] among others, as expected, given the role of a Th1 type response in fighting such conditions. They are also more affected by Th17 type conditions as seen in various models of arthritis. This is presumably because T-bet acts to negatively regulate the *IL17* locus through sequestration of the IL-17 activator Runx1.[179] By contrast, they are resistant to those inflammatory and auto-immune type conditions known to have a substantial Th1 type component such as type I diabetes (TID).[241] Expression of T-bet seems mostly limited to the immune system.[229] In addition to Th1 cells, T-bet expression can also be seen in B cells, NK cells, DCs, $CD8^+$ cells and Innate Lymphoid Cells (ILCs). T-bet is involved in promoting antibody class switching and loss of T-bet in B cells reduces class switching to IgG2a and prevents the B cell mediated

pathologies of a mouse model of lupus.[242] T-bet expressing DCs promote the development of inflammation in a mouse model of rheumatoid arthritis which is likely, in part, due to T-bet's positive regulation of IL-1α and MIP-1α in these cells.[243] By contrast, T-bet negatively regulates TNFα production in DCs. Mice lacking an adaptive immune system through deletion of Rag2 and also lacking T-bet can develop spontaneous colitis. Colonic DCs in these TRUC (for **T**-bet and **R**ag deficient **U**lcerative **C**olitis) mice hyperproduce TNFα which is necessary for development of the disease.[244] Development of colitis in the TRUC model also requires colonic ILCs. Wild-type ILCs can express T-bet which represses expression of IL7RA and IL-17A. The higher production of these two proteins contributes to the development of colitis in the TRUC model.[245]

As previously mentioned, loss of T-bet in NK cells affects their development and it also reduces their ability to produce IFN-γ and to lyse target cells. The loss of T-bet in $CD8^+$ cells affects their cytolytic ability and IFN-γ production downstream of specific antigen stimulation.[246] However, the loss of T-bet is less profound in $CD8^+$ cells than in $CD4^+$ cells due to redundancy between T-bet and Eomes.[188][182]

T-bet interacts with a variety of different proteins in order to perform different functions. It can recruit histone modifiers such as JMJD3, a H3K27me3 and -me2 demethylase and SET7/9, a H3K4me2 methytransferase.[247][248] Since H3K27me3 is a repressive mark whereas H3K4me2 is an activating mark, this allows T-bet to both remove repressive marks and add activating marks at key genes. Furthermore, because methylation marks are fairly stable and inheritable, some of the effects of T-bet remain at the locus even if T-bet itself disengages.

In naïve $CD4^+$ cells, the *TBX21* locus is in a bivalent state and shows permissive H3K4me3 and repressive H3K27me3 methylation. H3K27me3 is lost on progression to Th1 commitment with accompanying increases in H3K4me3. In the Th2 lineage, repressive H3K27me3 is increased but H3K4me3 is not completely lost. In Th17 and Treg cells the locus resembles the naïve state of combined H3K4me3 and H3K27me3, suggesting that T-bet is poised and can be reexpressed in these lineages.[8] Indeed, T-bet expression is seen in a subset of Treg cells that have been primed by IFN-γ to suppress Th1 like inflammation. The expression of T-bet in these cells allows homing of the Tregs to sites of Th1 type inflammation through the expression of T-bet target CXCR3.[159] Coexpression of master regulators is not restricted to Treg cells. In addition to T-bet, Th1

cells also express GATA3 which binds to many of its target genes, including *IL4* and *STAT6* in the Th1 lineage.[158] Since IL-4 is not expressed in the Th1 lineage, this further highlights the way in which gene expression results from integration of multiple genomic events and not, for example, just the binding of one transcription factor. T-bet interacts physically with GATA3[249] and co-binding of T-bet and GATA3 in Th1 cells suggests that one role of T-bet in Th1 lineage commitment is the redistribution of the GATA3 protein.[234]

T-bet physically interacts with many different proteins. An interaction with the NF-κB subunit RelA allows T-bet to progressively inhibit IL-2 production in $CD4^+$ cells as they become Th1.[250] T-bet can also physically interact with Runx3 which is needed for optimal IFN-γ production in Th1 cells[251] and Runx1 to prevent activation of the Th17 master regulator RORγT[179]. At a functional level, cooperation between T-bet and other transcription factors is even broader. Of particular note to this project is the cooperation between T-bet and STAT4. Unlike the master regulators, the STAT proteins have a very broad range of targets and functions. However, in the context of the immune system, STAT1 and STAT4 are considered to cooperate in the Th1 lineage commitment while STAT6 is involved in Th2 lineage commitment. STAT4 is required for proper *IFNG* chromatin remodelling and IFN-γ expression downstream of IL-18 signalling[170] and for IL18R1 expression.[213]

1.2.4.2 GATA3

In contrast to T-bet, GATA3 has a key role in embryonic development and full knockout of GATA3 in mice is embryonic lethal.[252] In the mature organism, GATA3 is needed for T cell development[253] in addition to Th2 lineage commitment and maintenance. In humans, there are six members of the GATA family named for their recognition of the GATA motif, (A/T)GATA(A/G). In vertebrates, the GATA family contains two highly conserved zinc fingers called the C-finger and N-finger after the terminus to which they are closest.[254] It is the C finger that recognises the GATA motif, although the N finger can also bind to DNA independently of the C finger, with preference for GATC rather than GATA.[255] Although the N finger generally binds less strongly than the C finger, it can bind sufficiently to activate gene expression. However, binding is strengthened and stabilised when both fingers bind. Recent profiling of GATA3 binding has shown that the half sites to which these two fingers bind can either

be three or four bp apart,[6] highlighting the complexity of analysing transcription factor binding from DNA sequence. Both fingers have also been shown to be involved in protein-protein interactions which are thought vital not just for GATA function but for GATA targeting to appropriate sites amongst a high frequency of motif occurrences across the genome. X-ray crystal structures of the C finger of mouse GATA3 have been published which suggest that the mode of GATA3 binding depends on the DNA sequence and relative concentration of GATA3.[256] While single GATA motifs will bind a single C finger, double GATA motifs may bind either both the C and N fingers of one GATA3 molecule or the C terminal fingers of two GATA3 molecules depending on relative spacing of the motifs and availability of a second GATA3 protein to bind.

In humans, the GATA family can be divided into two subfamilies. GATA1, GATA2 and GATA3 are expressed and important in the haematopoietic system while GATA4, GATA5 and GATA6 are mainly expressed in the heart, liver and small muscle cells. However, there are exceptions to these generalisations. GATA3, for example is also expressed in skin, prostrate and kidney. Consistent with their key roles on development, mutations in five of the six GATA proteins either cause or are associated with disease. The one exception is GATA5. Haploinsufficiency in GATA3 causes Hypoparathyroidism, sensorineural Deafness, and Renal disease (HDR), also known as Barakat syndrome.[257] The NHGRI GWAS catalogue reports a SNP in the *GATA3* locus as associated with Hodgkin's lymphoma and also reports disease associations for SNPs at *GATA2*, *GATA4* and *GATA6*.

GATA3 was originally described in chickens as NF-E1c.[258] Of note, its expression was seen in cells of lymphoid lineage in contrast to GATA1 (then called NF-E1) and GATA2 (then called NF-E1b). This lead a later study, examining the role of the human and mouse homologue of this gene, to find abundant expression of GATA3 in human T cells and a role for GATA3 in activating transcription of the human T Cell Receptor delta gene[259]. The role of GATA3 as the master regulator of the Th2 lineage, needed for expression of the Th2 cytokines, was later demonstrated by Zheng *et al* in 1997.[215]

The role of GATA3 in the immune system has been investigated using conditional knockouts and complementation systems. Such studies have shown that GATA3 is required for T cell development, in addition to its later role in Th2 lineage commitment.[253] In terms of its role in Th2 lineage commitment, GATA3

binds to the Th2 cytokine locus as discussed above. GATA3 can also autoactivate its own locus independently of STAT6 and IL-4.[260] Like T-bet, GATA3 interacts both physically and functionally with a number of different proteins. For example, GATA3 physically interacts with Runx3 and this prevents GATA3 from binding to and activating its transcriptional targets.[261] GATA3 can also bind to the promoter of the Treg master regulator *FOXP3* and inhibits the expression of this factor and commitment to the Treg lineage.[262]

Given its role in Th2 lineage commitment, it is unsurprising that GATA3 has been implicated in pathologies associated with a Th2 type response. Higher levels of GATA3 are seen in the airways of asthmatic patients[263] and in nasal biopsies from patients with allergic rhinitis.[264] However high levels of GATA3 are also associated with breast cancer survival[265,266] consistent with a role of Th2 cells in cancer immunosurveillance.

1.2.5 Pathological Consequences of Immune System Dysregulation

As discussed above, the immune system is finely balanced and the direction of an immune response results from integration of many signals and events. Mechanisms exist to keep the immune system balanced. However, the various feedback mechanisms that allow strong, robust and recallable responses to pathogens can also, if wrongly triggered, allow such responses to host tissue or non pathogenic foreign substances such as drugs or food. This results in autoimmune or inflammatory conditions. The most common conditions are complex in nature, resulting from a range of genetic and environmental factors and the interplay between the two. The relative contribution of genes and environment varies across the spectrum of common complex diseases and across the people that suffer from them. To highlight some of these points, some of the genetic and environmental factors and the role of the immune system in the complex conditions of Type I Diabetes (T1D), coeliac disease, obesity and the inflammatory bowel diseases ulcerative colitis and Crohn's are discussed briefly below.

1.2.5.1 Type I Diabetes

Diabetes usually results from an impaired ability to produce insulin in the body (type I) or to respond to the insulin that is produced (type II). In both cases, the body is unable to upregulate glucose metabolism and physical symptoms include high levels of thirst and tiredness. As its alternative name, juvenile diabetes, suggests T1D commonly occurs in younger people. Until the medical use of insulin began in 1922, T1D was a fatal condition. Although no fully effective treatment exists, TID can now be controlled by regular injections of insulin combined with healthy diet and other behaviours. More permanent treatments for T1D including islet transplantation have been developed.[267] However, technical issues remain and such treatments are not currently in widespread use.

The presence of autoantibodies and autoreactive T cells make T1D an exemplar of autoimmunity. Genetically, the largest risk factor is the haplotype at the Human Leukocyte Antigen (HLA) Class II locus. At least 90% of patients have class II HLA genes of type either DRB1*04-DQB1*0302 or DRB1*03.[268] These presumably allow good presentation of autoantigenic peptides such as peptides from preproinsulin and glutamic acid decarboxylase to $CD4^+$ cells. These $CD4^+$ then provide help to $CD8^+$ cells in destroying the pancreatic tissue from which those peptides emanate. However, this is clearly not the only risk factor. Although 90% of patients have the risk alleles, only 5% of people with the risk alleles go on to develop T1D.[67] The haplotype at the HLA class I locus, which allows peptide presentation directly to $CD8^+$ T cells, is also a risk factor.[269] Other risk alleles are found at the insulin gene itself and various other immune-gene loci such as *CD25* (IL2Rα)[270] and *IL10*.[271] Rare protective variants have been found at the *IFIH1* locus highlighting the importance of rare variant hunting in addition to GWAS.[272] Some SNPs that predict increased risk in T1D predict decreased risk in other conditions. For example, SNPs in the *IL18RAP* and *TAGAP* loci are associated with decreased risk of TID but increased risk of coeliac disease.[273] In general, there is much overlap between disease associated loci for the different autoimmune and inflammatory conditions. For T1D, many of these overlapping loci have been annotated on T1Dbase (http://www.t1dbase.org/page/-Welcome/display). Overlapping loci include *RGS1*, which is associated with T1D and also coeliac disease and multiple sclerosis. SNPs in *IL10* are associated with ulcerative colitis and coeliac disease in addition to T1D and SNPs in *STAT4* are associated with T1D and also rheumatoid arthritis, systemic lupus

erythematosus and, again, coeliac disease.

In terms of understanding the underlying immunology, GWAS have helped by confirming the importance of the MHC region in disease susceptibility and suggesting other avenues of investigation in understanding the disease. Such avenues have been reviewed recently[268] and include a potential role for B cells, as highlighted by the discovery of associated SNPs near *BACH2*, a gene involved in B cell gene regulation.

The incidence of T1D has been steadily rising over the past few decades pointing to a role for environmental factors in disease mechanism. Type 1 diabetes was initially thought to be triggered by infection on a genetically susceptible background. However, more recent work has shown that, while some pathogens can trigger the immune response, the relative lack of immune challenge in the modern world can also increase the risk of T1D. Evidence for other environmental risk factors has also been found including early introduction to solid food, childhood obesity and birth by caesarean section.[274] Although much work remains to be done to confirm and evaluate the roles of such factors, these results highlight the complex interplay between different aspects of the immune system and different environmental factors.

1.2.5.2 Ulcerative Colitis and Crohn's Disease

Both ulcerative colitis and Crohn's disease have an aberrant immune and inflammatory component and are thought to result from an immune reaction to microbial products in the gut.[67] Traditionally, these diseases have been described as having distinct but overlapping pathologies. Macroscopically, ulcerative colitis is characterised by ulcers in the colon and it has been categorised as a Th2 type condition. Published genetic susceptibility loci include genes important for epithelial barrier function such as *HNF4A*.[275] By contrast, Crohn's disease is characterised by lesions throughout the small and large bowel and has generally been considered to be Th1 driven. GWAS susceptibility loci include *NOD2* (also know as CARD15)[276,277] and *ATG16*[71] which has implicated the innate immune system and autophagy in disease mechanism. However, both diseases share multiple susceptibility genes, notably including some involved in the Th17 lineage.[278] Furthermore, a recent large meta-anlysis of the two diseases has shown that, at the genetic level, the diseases are very similar.

Most susceptibility loci are, in fact, shared between the two diseases. It is now thought that the same biological pathways are involved, to a greater or lesser extent in both conditions.[279]

The role of the immune system in both diseases has not only been highlighted by genetic associations with immune related genes but also by various mouse models. Transfer of naïve $CD4^+$ cells into an immunodeficient host allows rapid proliferation of the transferred cells and the mice lose weight and develop inflammation in the gut,[280] although not if the transferred cells are deficient in T-bet.[281] Mice lacking an adaptive immune system through deletion of the *RAG2* gene but also lacking the T-bet transcription factor can develop bowel diseases resembling ulcerative colitis. Disease from these TRUC (T-bet$^{-/-}$ $\times$ Rag2$^{-/-}$) mice is transmissible to immunocompetent hosts and onset can be prevented by antibiotic treatment which implicates bacteria in the disease initiation.[244] One bacterial species that can trigger colitis under these genetic conditions has now been identified as *Helicobacter typhlonius*. Although there may be others, this species was absent in a rederived T-bet$^{-/-}$, Rag2$^{-/-}$ double knockout colony in which the mice did not develop spontaneous colitis.[245] Furthermore, administration of *H. typhlonius* to members of the rederived colony induced colitis. By contrast, administration of *H. typhlonius* to Rag2$^{-/-}$ and T-bet sufficient mice did not induce the same levels of colitis. This shows the importance of the interplay between gut microbiota composition and a genetically susceptible background for disease development.

Another example of this was described by Cadwell *et al*[282] who showed that administration of a persistent strain of murine norovirus can induce paneth cell abnormalities in mice with mutant ATG16L1. These abnormalities are not seen in wild-type mice infected with the same strain of norovirus or in ATG16L1 mutant mice without infection. Mutant mice carrying the infection displayed villus blunting on administration of low doses of dextran sodium sulphate which was not seen in infected wild-type or non-infected mutant mice. Thus, a genetic factor had allowed an invading pathogen to promote phenotypic abnormalities in the gut which then worsened outcome to a further environmental challenge. Further examples of this interplay have been reviewed by Virgin and Todd.[67]

1.2.5.3 Coeliac Disease

In coeliac disease, an immune reaction is mounted to a foreign peptide but one that is non pathogenic and present in food, specifically gluten. Although it is therefore not strictly autoimmune, coeliac disease does share similarities with T1D. Antibodies are found in the blood against the offending peptides and to transglutaminase the enzyme that modifies them.[283] Furthermore, the HLA class II region has been found to be the strongest genetic susceptibility locus confirming the importance of the immune system in the disease[284]. Other immune related susceptibility loci shared between Coeliac disease and T1D include *CTLA4*, *TAGAP* and *IL18RAP* although SNPs in *IL18RAP* and *TAGAP* were shown to have opposite effects in susceptibility to or protection from T1D and coeliac disease.[273] The overlap in regions found by GWAS between coeliac disease and T1D is not surprising as the two diseases have a tendency to co-occur.[285] Other disease susceptibility loci shared between coeliac disease and other conditions include *PTPN2* (shared with Crohn's disease and T1D), *ETS1* (shared with systemic lupus erythematosus) and *LPP* (shared with vitiligo).[286]

Studies of coeliac disease in monozygotic twins show a concordance of approximately 75%,[287] suggesting the disease is highly heritable with a strong genetic component. However, symptoms are also entirely dependent on environment, specifically the presence or absence of gluten in the diet, highlighting the importance of the gene-environment interaction in disease progression.

Typical symptoms of coeliac disease include diarrhoea, bloating and abdominal pain. In the UK, a diagnosis of coeliac disease follows a blood test for the appropriate antibodies and then a confirmatory biopsy of the small intestine. Because of the highly heritable nature of coeliac disease, such testing is performed on close family members of anyone who becomes diagnosed with coeliac disease in addition to people with the typical symptoms. Testing is also performed on type 1 diabetics. Coeliac disease is recorded to affect approximately 1% of the population. However, there is some thought that this could be an underestimate due to underreporting of milder cases.

1.2.5.4 Obesity

Obesity is usually measured on the basis of a individual's Body Mass Index, with Body Mass Index of between 25 and 30 defined as overweight and of over 30 defined as obese. Obesity is associated with increased risk of type 2 diabetes, cardiovascular disease and some types of cancer.[288]

Obesity rates started noticeably rising in high income countries in the 1970s. As such, the rise has occurred over too brief a timescale to be fully attributed to genetic changes. The increase has coincided with more food availability and less requirement for physical activity (the so-called obesogenic environment). Interventions which tackle this environment, including low fat diet programmes and reductions in advertising of unhealthy food to children have moderate to strong evidence of effectiveness.[289] Clearly then, the environment plays a huge role in obesity. However, obesity is a heterogeneous condition and not all individuals in the obesogenic environment are obese suggesting other factors are important. At the extreme end, rare cases of obesity have a very strong genetic component. Leptin regulates the feeding response and disruption of the *LEP* gene (also known as Ob) results in overeating and increase in body weight as described in mice lacking the leptin gene (Ob/Ob mice).[290,291] In humans, individuals that are homozygous for mutations in the *LEP* gene also develop severe obesity associated with increased demand for feeding as a result of loss of leptin secretion and function.[292] Another condition, Prader-Wili syndrome, occurs because a section of the paternally derived chromosome 15 is not expressed, often due to a 4Mbp deletion in the region[293]. Prader-Wili syndrome is characterised by multiple behavioural and developmental issues but also food hoarding and obesity.

More commonly, GWAS have found multiple regions associated with obesity. The first GWAS hits, in the *FTO* gene, associated with an increased body mass of 3kg in the homozygous risk allele group[294] but other SNPs have been found near genes such as *MCAR4* and *NPC1*.[295] Some risk loci have been found to overlap between studies of extreme obesity and studies of overall Body Mass Index suggesting that studying the genetics of extreme obesity can yield important insight into less extreme and more common forms of the condition.[296] These loci include both *FTO* and *MCAR4* as well as *TMEM18*, *NRXN3*, *BDNF* and *FAIM2*. In addition to its association with Body Mass Index, *FTO* is also associated with waist circumference and fat mass.

The immune system has also been implicated in obesity, as has a role for the gut microbiota. Weight loss is associated with changing microbiota composition and germ free mice gavaged with microbiota from mice fed on a high fat 'Western' diet gain more weight than germ free mice gavaged with microbiota from lean controls.[297] Mice deficient for the innate immune molecule TLR5 have insulin resistance, increased adiposity and a different microbiota from wild-type controls. Microbiota from these mice can confer obesity and insulin resistance when transferred to wild-type germ free mice.[298] Furthermore, recent publications have shown the importance of adipose infiltration by immune cells such as T cells and macrophages.[299]

1.3 Hypothesis

In this project, we hypothesise that some binding sites for the transcription factors T-bet and GATA3 will contain genetic variation in the form of SNPs, which will alter the binding of the transcription factor to the DNA at that site and consequentially alter the regulation of gene expression. Because T-bet and GATA3 have a central role in the development of an immune response we hypothesise that such variation and its downstream effect on genomic regulation will alter disease risk. As such, we hypothesise that this variation will be found amongst the significantly disease-associated SNPs found in GWAS. Examining GWAS SNPs with respect to genomic features such as the presence of a transcription factor binding site will provide a way to causally link genetic variation and disease mechanism.

1.4 Aims

Specific aims of this project are:

1. To examine whether disease-associated genetic variation in the form of SNPs found from GWAS can be analysed and annotated with respect to alterations in transcription factor binding on a broad scale across multiple diseases and traits.

2. To apply the ideas from large scale projects such as ENCODE to cells and

transcription factors that are functionally relevant to disease mechanisms.

3. To establish and evaluate components of an *in vitro* pipeline for moving from statistical association from wide-scale genomic data to an *in vitro* mechanism and *in vivo* pathological consequences thus evaluating best way to move from *in silico* analysis to disease effects.

4. To further examine the role of the T-bet and GATA3 transcription factors in disease mechanisms.

2

Materials & Methods

2.1 *In Silico* Analysis

2.1.1 Determination and Annotation of SNPs in Binding Sites for T-bet and GATA3

Data on binding sites for the transcription factors T-bet and GATA3 had previously been obtained by ChIP-Seq.[234] The authors had performed the ChIP-Seq for T-bet with a polyclonal antibody; ChIP-Seq for Gata3 had been with a monoclonal antibody (D-16, Santa Cruz). Peak summits had been called from tag coordinate bed files using MACS (Model-based Analysis of ChIP-Seq), normalising to whole cell extract from Th1 or Th2 cells as appropriate, with a p-value cut-off of 1×10^{-6}. Binding sites were then defined as 100bp either side of each summit. The GWAS catalogue was downloaded from the National Human Genome Research Institute (http://www.genome.gov/gwastudies/index.cfm) on November 4th 2011. Some entries did not contain information on the chromosome on which the SNP was found and this was added in using dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/). In this process, two SNPs were deleted as they were called suspect or unverified by dbSNP. SNPs on the X and Y chromosomes were also deleted for the systematic analysis. Data was analysed using the bioconductor snpMatrix package (now updated to snpStats) within an in-house R based script (see appendix).[96,97,300] Chromosomal locations of SNPs and identification of SNPs in high LD (r^2> 0.8 with a SNP from the GWAS catalogue) were obtained from HapMap3.[74]

Tag coordinate bed files for H3K4me1, H3K4me3 and H3K27me3 ChIP-Seq were obtained from previously published data.[10] Peaks were called using the MACS algorithm on the Galaxy/Cistrome platform (http://cistrome.org/ap/root).[301] We used a p-value cut-off of 1×10^{-6}. We were unable to normalise the data to a control dataset as there was none available. Therefore, normalisation was to local genomic background within the sample (using lambda = TRUE as a setting for MACS.) The MACS algorithm has been shown to function with reasonable reliability under these conditions.[302] We set the algorithm to deal with duplicate tags by modelling the genome on 'Auto by Binomial' and the data was produced and analysed on the Hg18/b36 annotation of the human genome. All other settings were kept as default. Peaks were taken as called and overlapped with binding sites for T-bet or GATA3 using the main Galaxy platform (http://-galaxyproject.org/).[98–100] Overlap was performed by selecting the 'Operate on Genomic Intervals' and then the 'Intersect' options within Galaxy. Intersection returned full length of any transcription factor binding site that overlapped a site of histone modification by at least 1bp. Data on DNase hypersensitivity was obtained from ENCODE.[303,304] (GEO Accession GSM736592 (for Th1 cells) and GSM736502 (for Th2 cells)). The authors had isolated Th1 and Th2 cells from peripheral blood using bead based $CD4^+$ T cell selection followed by Fluorescence Activated Cell Sorting (FACS) based on expression of surface markers. Surface markers for Th1 had been $CD45RO^+$, $CCR6^-$, $CCR4^-$, $CXCR3^+$. Surface markers for Th2 had been $CD45RO^+$, $CCR6^-$, $CCR4^+$, $CXCR3^-$). Cells had then been subjected to DNase-Seq. Detailed protocols for data generation can be found on the ENCODE website. Narrowpeak files on genome build Hg18/b36 were uploaded onto Galaxy. DNase hypersensitive peaks on autosomal chromosomes in Th1 cells were intersected with binding sites for T-bet and for GATA3 in Th1 cells. DNase hypersensitive sites on autosomal chromosomes in Th2 cells were intersected with binding sites for GATA3 in Th2 cells. Intersections were performed as for histone modifications. Transcription factor binding site peaks that intersected a DNase hypersensitive site were analysed for the presence of SNPs as above.

Locations of all hit SNPs were downloaded from dbSNP using dbSNP Batch query (http://www.ncbi.nlm.nih.gov/SNP/batchquery.html). Results were returned as a chromosome report which gave locations of all SNPs mapped to genome build Hg19/ 37. We used Galaxy to obtain 100bp of sequence either side of each SNP in FASTA format. These were submitted to MEME (Multiple

Em for Motif Elicitation) through the web interface to search for other transcription factor binding motifs. We also looked for published motifs for T-bet and GATA3 using the `dreg` tool from European Molecular Biology Open Software Suite (EMBOSS) and using the FIMO (Find Individual Motif Occurrences) tool through the web interface.[305,306] In each case the FASTA sequences were submitted as obtained and also with each SNP manually altered to the opposite allele using dbSNP for reference. For the `dreg` tool, we searched 20bp either side of the SNP for the following regular expressions:

For T-bet: (tt[atcg][atcg]cac) | (cacac) | (gtgtg) | (gtg[atcg][atcg]aa)

For GATA3: (gata) | (tatc)

For the FIMO tool, we used the motifs generated for T-bet and GATA3 as generated by Kanhere *et al*[234] and ran these against the entire 200bp FASTA sequence as obtained from Galaxy.

2.1.2 Testing for Transcription Factor Co-binding

Potential binding of other transcription factors in the regions around hit-SNPs was examined using MEME[306] and TRAP (Transcription Factor Affinity Prediction).[307] For MEME analysis, sequences were uploaded onto web interface and settings used for analysis were as follows:

Distribution of occurrences of a single motif across sequences = Any number

Minimum motif width = 3

Maximum motif width = 50

Maximum number of motifs to find = 10

All optional settings were left blank.

Results included many motifs that were not well conserved across multiple input sequences. Manual examination revealed one motif that was well conserved across multiple sequences and this was found to match the Runx motif by manual comparison to the JASPAR database.

For TRAP analysis, sequences were uploaded onto the web interface and analysed for potential transcription factor binding events against the JASPAR vertebrates database with background model of human promoters. TRAP ranks

transcription factor binding based on biophysical measurements of affinity between DNA and transcription factor rather than using a PWM. The programme returns a p-value reflecting the affinity measurement and a corrected p-value which accounts for multiple testing over many sequences. Control sequences, of the same length as the hit-SNP sequences, but that were chosen as random sequences from the human genome, were also analysed. The highest corrected p-value reached for the control sequences was taken as the cut-off for background noise in the hit-SNP analysis and only transcription factors scoring more highly than this value are listed in the results.

2.1.3 Testing for Enrichment of Inflammatory Conditions

A list of all unique traits in the GWAS catalogue was compiled and each trait was scored independently as definitely immune mediated (Y), possibly immune mediated (?) or definitely non immune mediated (N) by three clinicians. Scores were then aggregated. Final scores were assigned to traits using the system shown in table 2.1 to give a list of conditions on which there was consensus of definite immune involvement (Y) and a list of conditions in which there was fair support for the idea of some immune involvement (?). These lists can be found in the appendix. Of note, we were stricter with our definition of an immune mediated trait than of a non immune mediated trait. This was because it was the lists of immune mediated traits that we intended to use in our subsequent analysis. Hence, traits were only called strictly immune mediated if they were scored as strictly immune mediated by all three clinicians. By contrast, traits were classified as non immune mediated if they were were scored as non immune mediated by all three clinicians or if they were scored as non immune mediated by two clinicians and possibly immune mediated by a third.

Once each trait had been classified, SNPs were classified as strictly immune mediated if they had only been found in GWAS for traits classified as Y, non-immune mediated if they had only been found in GWAS for traits classified as N and loosely immune mediated otherwise.

Classification 1	Classification 2	Classification 3	Overall Classification
Y	Y	Y	Y
?	Y	Y	?
?	?	Y	?
N	N	N	N
?	N	N	N
?	?	N	?
Y	N	N	?
Y	?	N	?

Table 2.1: Classifying traits. Traits in the NHGRI GWAS catalogue were classified as to immune status by three clinicians (classification 1-3). Overall classification was assigned based on independent classifications as shown.

2.1.4 Other Transcription Factors

Data for NF-κB binding sites was obtained from Kasowski *et al*[31] (Geo Accession = GSE19486). The authors had performed ChIP-Seq on cross-linked cells from different lymphoblastoid lines generated by the HapMap project. Data had been processed through the Illumina pipeline and peaks called through the PeakSeq pipeline. We combined all files for NF-κB binding in all the different lines tested and merged intersecting peaks using Galaxy. We then analysed the resulting dataset. Data for STAT4 binding was obtained from Liao *et al*[308] (Geo accession = GSE27158). The authors had sorted $CD4^+$, $CD45RA^+$ cells from human buffy coats, pre-activated with αCD3 and αCD28 and stimulated with IL-12 prior to cross-linking and ChIP-Seq. Reads had been mapped to the genome using the Illumina pipeline. Files were available as tag coordinate bed files which we ran through the MACS analysis as with the methylation data but using a p-value cut-off of 1×10^{-5}. Data for ER-α binding was obtained from Stender *et al*.[309] Authors had performed ChIP-Seq in human breast cancer cells. Processing of the peaks had been performed by authors using ELAND and peak identification was by authors using HOMER.

2.2 Mice

Wild-type BALB/c and C57BL/6 mice were obtained from Charles River Laboratories International (Margate, UK). T-$bet^{-/-}$ mice on both BALB/c (for adoptive transfer experiments) and C57BL/6 (for all other experiments) backgrounds

were from Taconic (Ejby, Denmark). IFN-$\gamma^{-/-}$ mice were on BALB/c background and were from Charles River Laboratories. Rag2$^{-/-}$ mice on BALB/c background were from Harlan Laboratories (Indianapolis, USA) and Rag1$^{-/-}$ mice on C57BL/6 background were a kind gift from R. Noelle. Mice were kept at Charles River Laboratories and then transferred to in-house facilities at Kings College London for experiments. Mice were kept in Specific Pathogen Free conditions and in accordance with home office protocols under a laboratory wide project license (reference number PPL70/6792). Mice were female except for three recipient mice in the coeliac experiment which were male due to mouse availability. Mice were between 4 and 12 weeks of age at start of experiment and knockout and wild-type control animals were age-matched as far as technically possible.

2.3 Cell Culture

2.3.1 Reagents

Cells were cultured in High Glucose (4.5g/l) Dulbecco's Modified Eagle Medium (DMEM) (PAA) or Roswell Park Memorial Institute 1640 (RPMI) medium (PAA). Both were purchased without L-Glutamine but then supplemented with 50 units/ml Penicillin, 50 μg/ml Streptomycin (obtained as combined Pen-Strep from Gibco), 10 mM HEPES buffer solution (Fisher Scientific), 1 mM Sodium Pyruvate (Gibco), 1 $\times$ Minimum Essential Medium-Non Essential Amino Acids (Gibco), 2 mM L-Glutamine (Gibco) and 10% Foetal Bovine Serum (PAA). For culture of mouse cells, 50 μM 2-Mercaptoethanol (Gibco) was also added. All cells were maintained at 37°C in 5% CO_2. Centrifugation of cells in preparation and analysis steps was at 652g for five minutes at either room temperature (when preparing cells for culture) or 4°C (when preparing cells for analysis) unless otherwise stated.

Phosphate Buffered Saline (PBS) without Ca^{2+} and Mg^{2+} was made in house from tablets (Oxoid) dissolved in water and autoclaved as appropriate except for coeliac model experiments where premade PBS (Gibco) was used.

2.3.2 Isolation and Culture of Cells from Mice

Mouse $CD4^+$ for culture and for *in vivo* models were obtained from spleen. Spleens were forced through mesh (SEFAR, reference number 3053-1000-676-00) to produce a single cell suspension. Red blood cells were lysed with ACK Lysis Buffer (0.15 M NH_4Cl, 1 mM $KHCO_3$, 0.1 mM Na_2EDTA (all from Sigma)). $CD4^+$ cells were enriched from total population using CD4 (L3T4) MACS beads (Miltenyi Biotec) according to manufacturers protocol. Naïve or effector/memory $CD4^+$ cells were then sorted by FACS based on markers for CD4, CD25, CD62L and CD44 as detailed in text and in table 2.2. Purity in post-sort analysis was typically 98% of live cells.

For cell culture experiments cells were activated in tissue culture treated 48 well plates (Corning) that had been pre-coated at 4°C overnight with 2 μg/ml αCD3 (clone 145-2C11 from BioXCell) and 2 μg/ml αCD28 (clone 37.51 from BioXCell) in PBS. To skew cells towards a Th1 phenotype, cells were plated at 1×10^6 per well in 0.5 ml RPMI supplemented as detailed above plus 20 ng/ml IL-12 (eBioscience, reference number 14-8121-62), 5 μg/ml αIL-4 (clone 11B11 from BioXCell) and 20 ng/ml IL-2 (BioLegend, reference number 589102). For Th2 conditions, cells were cultured in RPMI plus 20 ng/ml IL-4 (eBioscience, reference number 14-8041-62), 20 ng/ml αIFN-γ (clone XMG1.2 from BioXCell) and 20 ng/ml IL-2 (as for Th1). Cells were left for 36 hours and then a further 0.5 ml RPMI containing the same cytokine and antibody cocktail was added per well. Cells were removed from activation into new 48 well plates approximately 72 hours after initial plating depending on activation state and split as appropriate. RPMI plus 20 ng/ml IL-2 was added whenever cells were split. Times for IL18R1 expression timecourse were measured from time of removal from activation.

For *in vivo* experiments cells were resuspended in 250 μl PBS and injected intraperitoneally as detailed below.

For analysis, spleens were harvested as for cell culture and *in vivo* models but either surface stained after red blood cell lysis with appropriate antibodies (as detailed in table 2.2) or stimulated with Phorbol 12-myristate 13-acetate (PMA) and Ionomycin prior to intracellular staining (as detailed below). Lymph nodes were also harvested for some experiments. Single cell suspensions were made and stained as for spleen except red blood cell lysis step which was not per-

formed.

2.3.3 Isolation and Culture of Cells from Human

Human $CD4^+$ cells were obtained from buffy coats that had been obtained through the National Blood Service or from whole blood from healthy volunteers under REC reference number 10/H0804/65. $CD4^+$ cells were isolated using RosetteSep Human CD4+ T cell Enrichment Cocktail (STEMCELL Technologies) according to manufacturers instructions. Briefly, blood was incubated with RosetteSep Mix and then layered over LSM 1077 Lymphocyte Separation Medium (PAA). Sample was then placed in centrifuge and run at 652g for 30 minutes at room temperature with acceleration set to half of maximum and with no break. $CD4^+$ cells migrated to a central layer in the sample from which they were removed by pasteur pipette and washed with PBS. $CD4^+$ cells were then activated in tissue culture treated 48 well plates that had been precoated at 4°C overnight with 2 μg/ml αCD3ϵ (clone UCHT1 from R&D systems) and 2 μg/ml αCD28 company (clone 37407 from R&D systems) in PBS.

To skew cells towards a Th1 phenotype, cells were plated at 1×10^6 per well in 1 ml supplemented RPMI plus 10 ng/ml IL-12 (R&D Systems, reference number 219-IL) and 10 μg/ml anti IL-4 (clone 34019 from R&D systems). After approximately 48 hours, cells were activated and were removed from activation into fresh 48 well plates. Cells were split as appropriate and wells were maintained at a volume of 1 ml by addition of RPMI containing 10 ng/ml IL-2 (BioLegend, reference number 589102). Cells were harvested to make lysate for OligoFlow after seven days of culture.

2.3.4 Cell Lines

YT cells were a kind gift from S. Johm, EL-4 cells were a kind gift from H. Stauss and Jurkat cells were a kind gift from G. Lombardi. YT cells were maintained in RPMI supplemented as described above. EL-4 and Jurkat cell lines were maintained in DMEM supplemented as described above. All cell lines were passaged every two to three days to a density of 1×10^6 per ml for YT cells, 0.5×10^6 for Jurkat cells and 1×10^5 for EL-4 cells.

2.4 Flow Cytometry

2.4.1 Reagents

Antibodies used for flow cytometry are detailed in (table 2.2.) All samples were acquired on Canto, LSR II or Fortessa machines (BD Biosciences) using Diva software. Multi-colour experiments were automatically compensated using One Comp eBeads (eBiosciences). Results were analysed in FlowJo version 7.6.3 (Tree Star Inc.)

Table 2.2: Antibodies. Antibodies used for flow cytometry experiments.

Reagent	Conjugation	Clone	Company	Catalogue Number	Dilution Used
Anti-Mouse CD62L	eFluor450	MEL-14	eBioscience	48-0621-82	1:100 (for analysis) or 1:20 (for sorting)
Anti-Human/Mouse CD44	FITC	IM7	eBioscience	11-0441-82	1:100 (for analysis) or 1:20 (for sorting)
Anti-Mouse CD4	PE-Cy7	GK1.5	eBioscience	25-0041-82	1:100 (for analysis) or 1:20 (for sorting)
Anti-Mouse CD25	PE	PC61.5	eBioscience	12-0251-81	1:100 (for analysis) or 1:20 (for sorting)
Anti-Mouse CD25	Alexa700	PC61.5	eBioscience	56-0251-82	1:200
Anti-Human/Mouse T-bet	PE	eBio4B10	eBioscience	12-5825-80	1:150
Anti-Mouse IFN-γ	eFluor450	XMG1.2	eBioscience	48-7311-82	1:150
Anti-Mouse IFN-γ	FITC	XMG1.2	eBioscience	11-7311-82	1:150
Anti-Mouse IL18Rα	Alexa647	BG/IL18RA	BioLegend	132903	1:40
Anti-Mouse CD4	APC-eFluor780	RM4-5	eBioscience	47-0042-82	1:200
Anti-Mouse CD4	FITC	GK1.5	eBioscience	11-0041-85	1:100
Anti-Mouse CD4	PE	H129.19	BD Biosciences	553652	1:200
Anti-Mouse CD3e	PE-Cy7	145-2C11	eBioscience	25-0031-82	1:200
Anti-Mouse CD4	APC	GK1.5	eBioscience	17-0041-82	1:20
Anti-Mouse CD25	APC-eFluor780	PC61.5	eBioscience	47-0251-80	1:20
Anti-Human/Mouse CD44	PE-Cy7	IM7	eBioscience	25-0441-82	1:200 (for analysis) or 1:20 (for sorting)
Anti-Mouse/Rat IL-17A	PE	eBio17B7	eBioscience	12-7177	1:150
Anti-Mouse IFN-γ	PE	XMG1.2	eBioscience	12-7311-82	1:100
Anti-Human GATA3	Alexa488	L50-823	BD Biosciences	560163	12.5 ng per Oligo-Flow Sample
Anti-Human/Mouse T-bet	eFlur660	eBio4B10	eBioscience	50-5825-80	0.1 μg per Oligo-Flow Sample
Anti-Human/Mouse T-bet	Alexa647	4B10	BioLegend	644803	Titrated to 0.25 μg per OligoFlow Sample (see text)

2.4.2 Surface and Intracellular Staining

Any samples to be analysed for intracellular cytokines were incubated in RPMI with 50 ng/ml PMA (Sigma) and 1 μM Ionomycin (Sigma) for two hours followed by addition of 3 μM monensin (Sigma) for a further 2.5 hours. Samples to be analysed were washed twice in PBS and incubated in 50 μl of 1:100 FcR Blocking Reagent (Miltenyi Biotec) in PBS for 10 minutes at 4°C. The surface stain consisted of 150 μl PBS containing surface antibodies (as detailed in text and in table 2.2) and 0.5 μl Live/Dead Fixable Aqua stain (Invitrogen, reference number L34957). Samples were incubated in the dark at room temperature for

20 minutes. Samples were washed twice in PBS and then either resuspended in PBS and acquired directly or fixed by resuspending in PBS and adding 37% paraformaldehyde dropwise to a concentration of 4% and a final volume of 1 ml. Cells were incubated in the dark for 10 minutes at room temperature then washed once in PBS. For surface stain experiments, cells were then resuspended in PBS and kept at 4°C in the dark until acquisition not more than 48 hours after fixation. For intracellular staining, after fixing, cells were washed once in Permeabilization Buffer (purchased as 10× concentrate from eBioscience and diluted to 1× with H_2O) and then resupended in 150 μl Permeabilization Buffer plus antibodies for intracellular components as detailed in text and in 2.2. Samples were incubated in the dark at room temperature for 20 minutes and then washed twice in Permeabilization Buffer. Centrifugation of cells prior to permeabilization was at 652g for five minutes at 4°C. All washes after permeabilization were at 887g for five minutes at 4°C. Samples were then resuspended in PBS and acquired.

2.5 Array data

Array data on IL18R1 and IL18RAP expression in T-bet$^{-/-}$, IFN-$\gamma^{-/-}$ double deficient cells transduced with T-bet or empty retrovirus was from Jenner *et al.*[158].

Data on IL18R1 and IL18RAP expression across T helper cells subsets in mice had been generated in house.[310,311] Briefly, naïve CD4$^+$ cells from Balb/c or C57BL/6 mice had been sorted by FACS based on expression of CD4$^+$, CD25$^-$, CD62L^{hi}, CD44^{low}. Cells for Th1, Th2 and Th17 had then been activated and skewed to appropriate conditions for seven days before cell harvest and RNA extraction for array. Balb/c cells for Th0 condition had been activated and cultured in medium plus IL-2 only for seven days. C57BL/6 cells for Thp condition were sorted directly into TRIzol reagent (Invitrogen) for RNA extraction.

2.6 *In Vivo* Disease Models

2.6.1 Coeliac Model

For the coeliac model, wild-type or T-bet $^{-/-}$ mice on C57BL/6 background were injected with 200 μg gliadin (Sigma) which had been sieved through a 100 μm cell strainer (Fisherbrand) to create a fine powder, suspended in 2 μl of 100% ethanol and then added to 100 μl Complete Freund's Adjuvant (CFA) (Sigma). Mice were then injected again 14 days later with 100 μg gliadin in 1 μl of 100 % ethanol in 100μl Incomplete Freund's Adjuvant (IFA) (Sigma). Spleens were harvested 28 days after first injection. Effector/memory CD4^{+} cells were then sorted by FACS as CD4^{+}, CD25^{-}, CD62L^{low} and CD44^{high}. Cells were resuspended in PBS and 4×10^5 effector/memory cells in 250 μl PBS were injected intraperitoneally into Rag1$^{-/-}$ mice on C57BL/6 background. All mice were maintained on RM3(E) pellets (Special Diet Services) which contains approximately 3-4% gliadin. Recipient mice were weighed once a week for eight weeks and then sacrificed. Spleen and lymph nodes were harvested for analysis by intracellular stain as detailed above and in results. Sections of proximal small bowel were harvested and fixed in 10% paraformaldehyde for histology staining or flash frozen in liquid nitrogen and stored at -80°C for RNA extraction and qPCR. Fixed histology sections were embedded in paraffin, cut and stained with Haematoxylin and Eosin as detailed by Powell *et al.*[245] For phenotyping the effector/memory cells used in the coeliac model, control mice were injected with 100 μl CFA only and then 100 μl IFA only 14 days later or 100 μl PBS followed by another 100 μl PBS 14 days later.

2.6.2 Adoptive Transfer Model of Inflammatory Bowel Disease

For the adoptive transfer model, naïve cells were prepared and sorted as above and 5×10^5 cells in 250 μl PBS were injected into Rag2$^{-/-}$ hosts on Balb/c background. Recipient mice were sacrificed after 2 weeks or 4 weeks and spleen and mesenteric lymph nodes were harvested and analysed by intracellular staining as detailed above and in results.

2.7 Oligonucleotide Pulldown

2.7.1 Buffers

Annealing buffer at 10 × concentration was 500 mM Tris pH8 (Fisher), 70 mM $MgCl_2$ (Fisher) and 10 mM Dithiothreitol (DTT, Sigma) and was diluted to 1 × with H_2O as appropriate. Hypotonic buffer was 20 mM HEPES pH8 (Sigma), 10 mM KCl (Sigma), 1 mM $MgCL_2$ (Fisher), 0.1% Triton-X (Sigma), 5% Glycerol, (Fisher) 1 mM DTT and 1 × Protease inhibitors (made from Complete Protease Inhibitor Cocktail tablets, Roche, reference number 11697498001). Hypertonic buffer was 20 mM HEPES pH8, 400 mM NaCl (Fisher), 1 mM EDTA (Sigma), 0.1% Triton (Sigma), 5% Glycerol, 1 mM DTT and 1 × protease inhibitors. Oligo buffer was 100 mM NaCl, 10 mM Tris pH8, 0.1 mM EDTA, 1 mM DTT, 5% Glycerol, 1 mg/ml Bovine Serum Albumin (BSA) Fraction V (PAA), 20 μg/ml dI/dC (Sigma, reference number P4929) and 1 × protease inhibitor. Sodium Dodecyl Salt (SDS) loading buffer was made at 5 × concentration as 312.5 mM Tris pH 6.8, 50% glycerol, 2.5% SDS (Fisher) and bromophenol blue (Sigma). Buffers were made fresh for each experiment except SDS loading buffer which was stored at 5 × and then diluted to 1 × and supplemented with 2% 2-Mercaptoethanol (Sigma) immediately prior to use.

2.7.2 Oligonucleotides Used

Table 2.3: Oligonucleotides. Forward and reverse oligonucleotides used for oligonucleotide pulldown and OligoFlow assays.

Oligo Name (=SNP)	Allele	Forward	Reverse
rs1006353	A	Biotin-TGGAGCACCCATCCA-CATAGGCCCCAATGATAAG-ATAGCGTTAACCACCAGC	GCTGGTGGTTAACGCTATC-TTATCATTGGGGCCTATGT-GGATGGGTGCTCCA
rs1006353	G	Biotin-TGGAGCACCCATCCA-CATAGGCCCCAGTGATAAG-ATAGCGTTAACCACCAGC	GCTGGTGGTTAACGCTATC-TTATCACTGGGGCCTATGT-GGATGGGTGCTCCA
rs1015290	A	Biotin-GTGAGGAGGTGGGA-GGAACATTCATGAGTGGGT-AACCCTGTGACTAATCCAG	CTGGATTAGTCACAGGGTT-ACCCACTCATGAATGTTCC-TCCCACCTCCTCAC

Oligo	Allele	Forward	Reverse
rs1015290	T	Biotin-GTGAGGAGGTGGGA-GGAACATTCATGTGTGGGT-AACCCTGTGACTAATCCAG	CTGGATTAGTCACAGGGTT-ACCCACACATGAATGTTCC-TCCCACCTCCTCAC
rs11135484	A	Biotin-AGAAACAACACCAA-CCTCACACCCACATAACAG-GATTAAAAGATAATGTGCA	TGCACATTATCTTTTAATCC-TGTTATGTGGGTGTGAGGT-TGGTGTTGTTTCT
rs11135484	G	Biotin-AGAAACAACACCAA-CCTCACACCCACGTAACAG-GATTAAAAGATAATGTGCA	TGCACATTATCTTTTAATCC-TGTTACGTGGGTGTGAGGT-TGGTGTTGTTTCT
rs13333528	C	Biotin-ACAGCTGCTTGGTGA-CGTTCATGTGGCTGGTTTCA-TTACCACCTAACTCCCA	TGGGAGTTAGGTGGTAATG-AAACCAGCCACATGAACG-TCACCAAGCAGCTGT
rs13333528	T	Biotin-ACAGCTGCTTGGTGA-CGTTCATGTGGTTGGTTTCA-TTACCACCTAACTCCCA	TGGGAGTTAGGTGGTAATG-AAACCAACCACATGAACG-TCACCAAGCAGCTGT
rs1420106	C	Biotin-AACTCTTCTCTGACC-AACCTCCTTGACGAGATAA-AGTTTGTGGTTGGGTTTA	TAAACCCAACCACAAACTT-TATCTCGTCAAGGAGGTTG-GTCAGAGAAGAGTT
rs1420106	T	Biotin-AACTCTTCTCTGACC-AACCTCCTTGATGAGATAA-AGTTTGTGGTTGGGTTTA	TAAACCCAACCACAAACTT-TATCTCATCAAGGAGGTTG-GTCAGAGAAGAGTT
rs1465321	A	Biotin-ACCTGCCATCCAGGA-GCTCCAAGGCTAGGGTTAA-CACTGAGGCCAACTGACC	GGTCAGTTGGCCTCAGTGT-TAACCCTAGCCTTGGAGCT-CCTGGATGGCAGGT
rs1465321	G	Biotin-ACCTGCCATCCAGGA-GCTCCAAGGCTGGGGTTAA-CACTGAGGCCAACTGACC	GGTCAGTTGGCCTCAGTGT-TAACCCCAGCCTTGGAGCT-CCTGGATGGCAGGT
rs2106346	A	Biotin-AACGTATGTTGACAG-ACTTTTCAAATAATCATCTT-TGTTATTACACCAAAAT	ATTTTGGTGTAATAACAAA-GATGATTATTTGAAAAGTC-TGTCAACATACGTT
rs2106346	C	Biotin-AACGTATGTTGACAG-ACTTTTCAAATCATCATCTT-TGTTATTACACCAAAAT	ATTTTGGTGTAATAACAAA-GATGATGATTTGAAAAGTC-TGTCAACATACGTT
rs2387397	C	Biotin-ATGCAAATAAGAAG-CTGTTTCAGTGTCTGCCCAT-CTGAGACGCTGACATAAA	TTTATGTCAGCGTCTCAGA-TGGGCAGACACTGAAACA-GCTTCTTATTTGCAT
rs2387397	G	Biotin-ATGCAAATAAGAAG-CTGTTTCAGTGTGTGCCCAT-CTGAGACGCTGACATAAA	TTTATGTCAGCGTCTCAGA-TGGGCACACACTGAAACA-GCTTCTTATTTGCAT
rs2703078	A	Biotin-GCTATGACCACTGGC-TCACATTTGCCACTTCAAG-CCTCCTACAAGAAACATA	TATGTTTCTTGTAGGAGGCT-TGAAGTGGCAAATGTGAGC-CAGTGGTCATAGC

Oligo	Allele	Forward	Reverse
rs2703078	G	Biotin-GCTATGACCACTGGC-TCACATTTGCCGCTTCAAG-CCTCCTACAAGAAACATA	TATGTTTCTTGTAGGAGGCT-TGAAGCGGCAAATGTGAG-CCAGTGGTCATAGC
rs2984920	A	Biotin-GATACACGTCACAG-CACACCAAGAAAAGGGGA-ACTTCCAGTGTCTGTGGTAA	TTACCACAGACACTGGAAG-TTCCCCTTTTCTTGGTGTGC-TGTGACGTGTATC
rs2984920	G	Biotin-GATACACGTCACAG-CACACCAAGAAAGGGGGA-ACTTCCAGTGTCTGTGGTAA	TTACCACAGACACTGGAAG-TTCCCCCTTTCTTGGTGTGC-TGTGACGTGTATC
rs3091310	C	Biotin-TAAAATCATTGTTCA-AATGAATGAATCAAGAGA-AGTTTAAACCACTTTGGAC	GTCCAAAGTGGTTTAAACT-TCTCTTGATTCATTCATTTG-AACAATGATTTTA
rs3091310	G	Biotin-TAAAATCATTGTTCA-AATGAATGAATGAAGAGA-AGTTTAAACCACTTTGGAC	GTCCAAAGTGGTTTAAACT-TCTCTTCATTCATTCATTTG-AACAATGATTTTA
rs5778	C	Biotin-ACTCTTCCAGCCTCC-CACATGATGGGCGGAAAA-AGGCAAAAGCCCAGATTAA	TTAATCTGGGCTTTTGCCTT-TTTCCGCCCATCATGTGGG-AGGCTGGAAGAGT
rs5778	T	Biotin-ACTCTTCCAGCCTCC-CACATGATGGGTGGAAAA-AGGCAAAAGCCCAGATTAA	TTAATCTGGGCTTTTGCCTT-TTTCCACCCATCATGTGGG-AGGCTGGAAGAGT
rs6784841	A	Biotin-CAGCTGCAGTGTATG-ACTATACATCAATGACTCA-CACGGTGGGGGATCCCTC	GAGGGATCCCCCACCGTGT-GAGTCATTGATGTATAGTC-ATACACTGCAGCTG
rs6784841	G	Biotin-CAGCTGCAGTGTATG-ACTATACATCAGTGACTCA-CACGGTGGGGGATCCCTC	GAGGGATCCCCCACCGTGT-GAGTCACTGATGTATAGTC-ATACACTGCAGCTG
rs743776	C	Biotin-ATCCAAACTCTCAGT-CTTGACCCACACGTCCTAC-AGGGACTGCCCCATCTTC	GAAGATGGGGCAGTCCCTG-TAGGACGTGTGGGTCAAGA-CTGAGAGTTTGGAT
rs743776	T	Biotin-ATCCAAACTCTCAGT-CTTGACCCACATGTCCTAC-AGGGACTGCCCCATCTTC	GAAGATGGGGCAGTCCCTG-TAGGACATGTGGGTCAAGA-CTGAGAGTTTGGAT
rs7441808	A	Biotin-TATTTACCTTAGGGA-CTCCTCTGGGTATGTGAAG-AATTCCCCTGTTTTGCTC	GAGCAAAACAGGGGAATT-CTTCACATACCCAGAGGAG-TCCCTAAGGTAAATA
rs7441808	G	Biotin-TATTTACCTTAGGGA-CTCCTCTGGGTGTGTGAAG-AATTCCCCTGTTTTGCTC	GAGCAAAACAGGGGAATT-CTTCACACACCCAGAGGA-GTCCCTAAGGTAAATA
rs8008961	C	Biotin-CATATAAGATTGCCT-CACATACTTGACTTTCATA-GAAGCGGAAGCCATTGAC	GTCAATGGCTTCCGCTTCT-ATGAAAGTCAAGTATGTGA-GGCAATCTTATATG

Oligo	Allele	Forward	Reverse
rs8008961	T	Biotin-CATATAAGATTGCCT-CACATACTTGATTTTCATA-GAAGCGGAAGCCATTGAC	GTCAATGGCTTCCGCTTCT-ATGAAAATCAAGTATGTGA-GGCAATCTTATATG
rs8062727	A	Biotin-CCACAGGGGAAAAA-TGTGTGGTTGCCACCACTT-CCTCTTATGGGGAAAGGAG	CTCCTTTCCCCATAAGAGG-AAGTGGTGGCAACCACAC-ATTTTTCCCCTGTGG
rs8062727	G	Biotin-CCACAGGGGAAAAA-TGTGTGGTTGCCGCCACTT-CCTCTTATGGGGAAAGGAG	CTCCTTTCCCCATAAGAGG-AAGTGGCGGCAACCACAC-ATTTTTCCCCTGTGG
T-bet Positive (T-bet +)	+	Biotin-TTAGCTAGTCGGCGC-TATAGTTTTCACACCTGATC-GTAGATCGCTAGCTAGTA	TACTAGCTAGCGATCTACG-ATCAGGTGTGAAAACTATA-GCGCCGACTAGCTAA
T-bet Negative (T-bet -)	-	Biotin-TTAGCTAGTCGGCGC-TATAGTTTTAAAACCTGAT-CGTAGATCGCTAGCTAGTA	TACTAGCTAGCGATCTACG-ATCAGGTTTTAAAACTATA-GCGCCGACTAGCTAA
GATA3 Positive V1 (Gata3+V1)	+	Biotin-TTAGCTAGTCGGCGC-TATAGGATCAGATAAGGTA-GATCGCTAGCTATAATTTG	CAAATTATAGCTAGCGATC-TACCTTATCTGATCCTATAG-CGCCGACTAGCTAA
GATA3 Negative V1 (Gata3-V1)	-	Biotin-TTAGCTAGTCGGCGC-TATAGGATCATAGAAGGTA-GATCGCTAGCTATAATTTG	CAAATTATAGCTAGCGATC-TACCTTCTATGATCCTATAG-CGCCGACTAGCTAA
GATA3 Positive V2 (Gata3+V2)	+	Biotin-ACCTGCCATCCAGGA-GCTCCAAGGATAGGGTTAA-CACTGAGGCCAACTGACC	GGTCAGTTGGCCTCAGTGT-TAACCCTATCCTTGGAGCT-CCTGGATGGCAGGT
GATA3 Negative V2 (Gata3-V2)	-	Biotin-ACCTGCCATCCAGGA-GCTCCAAGGCTGGGGTTAA-CACTGAGGCCAACTGACC	GGTCAGTTGGCCTCAGTGT-TAACCCCAGCCTTGGAGCT-CCTGGATGGCAGGT

2.7.3 Protocol

A sequence of 52bp around each allele of each SNP to be tested was obtained from dbSNP and this sequence was ordered with 5′ biotinylation (Integrated DNA Technologies). The reverse complement of each sequence was also ordered unmodified (see table 2.3). Oligonucleotides were resuspended to a concentration of 100 μM and stored at -20°C. Complementary sequences were annealed to produced double stranded oligonucleotides by mixing 10 μl of the forward oligonucleotide with 10 μl of its reverse complement, 5 μl of 10 $\times$ annealing buffer (described above) and 25 μl of DNase free water and incubating in a thermocycler for the following cycle: 94°C for 5 minutes, 65°C for 10 min-

utes, 25°C for 10 minutes and 4°C thereafter. Double stranded oligonucleotides were kept at 4°C until same-day use or stored at -20°C for long term use. For pulldown and Western blot, 20 μl of Streptavidin Agarose beads (Sigma) were used per sample. For OligoFlow, 50 μl of Sphero Streptavidin Polystyrene Particles (Spherotech, reference number SVP-100-4) were used per sample. Beads were washed twice in PBS and then once in 1$\times$ anneal buffer. For each wash, beads were centrifuged for three minutes at 7000g at room temperature, supernatant was removed and beads were resuspended in 500 μl. Beads were then incubated for one hour with oligonucleotides at 4°C. Beads were then washed twice in oligo buffer and resuspended in 450 μl oligo buffer. Lysate was prepared from between 10 and 30 million cells per sample. Cells used were YT or human $CD4^+$ cells cultured to Th1 *in vitro* for T-bet pulldowns or Jurkat cells for GATA3 pulldowns. Cells were washed twice in PBS and then resuspended in 1 ml hypotonic buffer and incubated on ice for five minutes. Lysed cells were pelleted at 4700g for five minutes, resuspended in 150 μl hypertonic buffer and vortexed vigorously for 12 minutes at 4°C. Debris was pelleted by centrifugation at 16000g for five minutes and 180 μl of supernatant containing nuclear extract was added to each sample. Samples were incubated on a rotor for one hour at 4°C. For Western blotting, samples were then washed three times in Oligo Buffer, resuspended in 25 μl SDS loading buffer and stored at -20°C until Western blot. For OligoFlow, antibody was added as appropriate (see text and table 2.2) and samples were covered and incubated for a further 1hr at 4°C. Samples were then transferred to ice and acquired on a Canto FACS machine.

2.8 Luciferase Assays

2.8.1 Primers

Region	Forward	Reverse
IFNG Promoter	AAAAAAGATCTGCCCTTCG-CATTCTTTCCTT	AAAAAAAGCTTCCCTTGG-TCCAAAGGACTTAA
Amplifying rs1465321	AAAAACTCGAGGAGATGA-AGAAATGAAGGTAGAGGG	AAAAAGCTAGCTGGATCT-CTATGACTTGTTGCTTAAA
Altering rs1465321	GTTAACCCTAGCCTTGGAG-CTCC	GGCTAGGGTTAACACTGA-GGCC

Table 2.4: Primers for cloning luciferase constructs.

2.8.2 Constructs

Firefly luciferase SNP testing constructs were based on PGL4.13 plasmid (Promega). All PCR programmes used in the construction of luciferase assay constructs were: 95.0°C for 5 minutes followed by 35 cycles of 95°C for 30s, appropriate annealing temperature for 30s and 72°C for 1 minute. PCR was then held at 72°C for 10 minutes and 4°C thereafter. Annealing temperature varied for different amplifications and is specified in each case below. IFNG promoter region from -456 to +68, relative to TSS was amplified using primers containing restriction sites for BglI (forward primer) or HindIII (reverse primer). PCR reaction mix was 1 μM IFNG promoter forward primer, 1 μM IFNG promoter reverse Primer, 1 μl template DNA and 22 μl Accuprime Pfx SuperMix (Invitrogen) in a final volume of 25 μl. Annealing temperature for PCR was 55°C. PCR product was gel purified and excised band was extracted with QIAquick Gel Extraction Kit (Qiagen). Resultant DNA product was digested with HindIII and BglII (both from New England Biolabs) and digested product was cleaned using QIAquick PCR Purification kit (Qiagen). PGL4.13 vector was digested with HindIII and BglII which removed unwanted SV40 promoter sequence. Resulting PGL4.13 backbone was gel purified as above. PGL4.13 backbone and IFNG promoter were ligated using T4 ligase (New England Biolabs). A 441 bp region around the SNP rs1465321 was amplified from genomic DNA using primers containing restriction sites for NheI and XhoI. PCR reaction mix was as for IFNG promoter region except primers were 'Amplifying rs1465321' primers detailed in table 2.4. Annealing temperature for amplification was 54°C. Once IFNG promoter had been successfully placed into PGL4.13 then vector was digested further with NheI and XhoI (both from New England Biolabs) and the region around rs1465321 was inserted using same method as for insertion of IFNG promoter region detailed above. Mutation of SNP to opposing allele was performed as previously described.[312] Briefly, two primers were designed against the region containing the rs1465321 SNP with base pair mismatch at the SNP itself. These 'Altering rs1465321' primers (as detailed in table 2.4) were paired with the 'Amplifying rs1465321' such that the forward 'Altering rs1465321' primer was used in reaction with the reverse 'Amplifying rs1465321' primer and vice versa to produce two half regions containing the opposite allele of the SNP. Remainder of PCR reaction mix was as before. Annealing temperature was 53°C. Products were gel purified as before and used as template for

a further round of PCR with 'Amplifying rs1465321' primers to produce full length DNA with the required mutation. Annealing temperature for final step was 53°C. DNA was digested and ligated into IFNG promoter containing vector as before. The final construct is shown in figure 2.1.

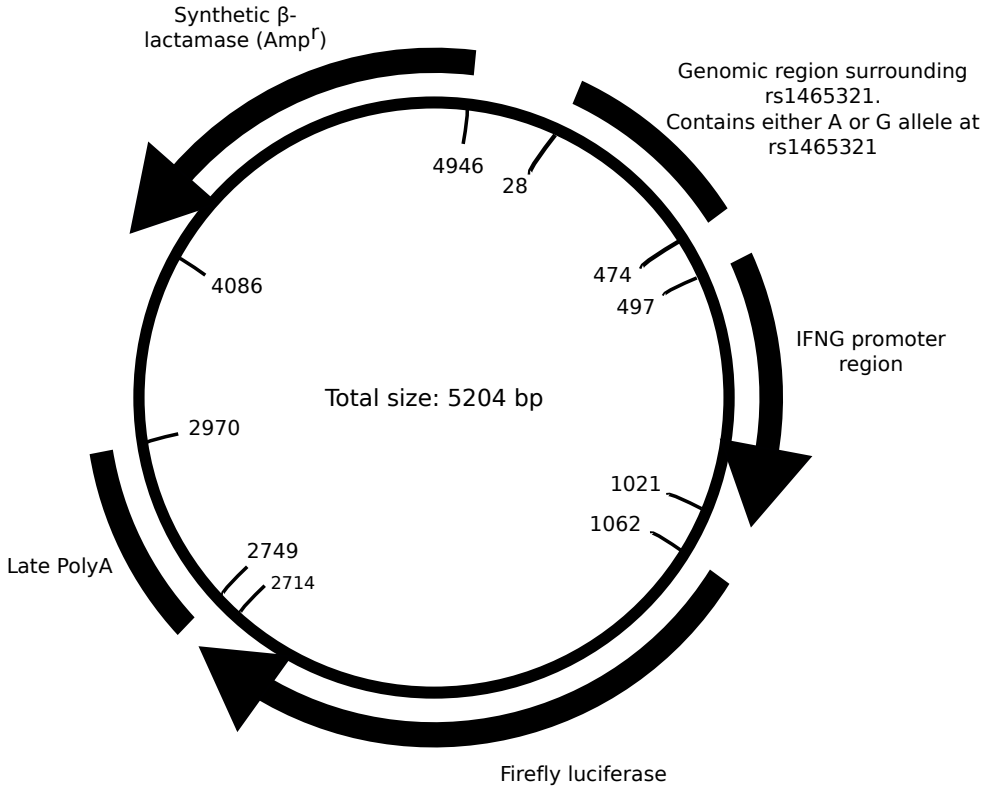

Figure 2.1: Luciferase construct for testing functional effect of altered T-bet binding at rs1465321. - Construct was based on PGL4.13 backbone with *IFNG* promoter and region surrounding rs1465321.

Human cDNA sequence for T-bet was digested from TOPO vector (Invitrogen) using XhoI and HindIII restriction enzymes (both New England Biolabs). Expression construct was based on pcDNA3.1 (+) (Invitrogen) and was also digested using XhoI and HindIII. T-bet sequence was ligated into construct as described for IFNG construct. All constructs were sequence checked before use and T-bet expression was tested by flow cytometry (fig. 2.2). Renilla luciferase construct (Promega) was used unmodified.

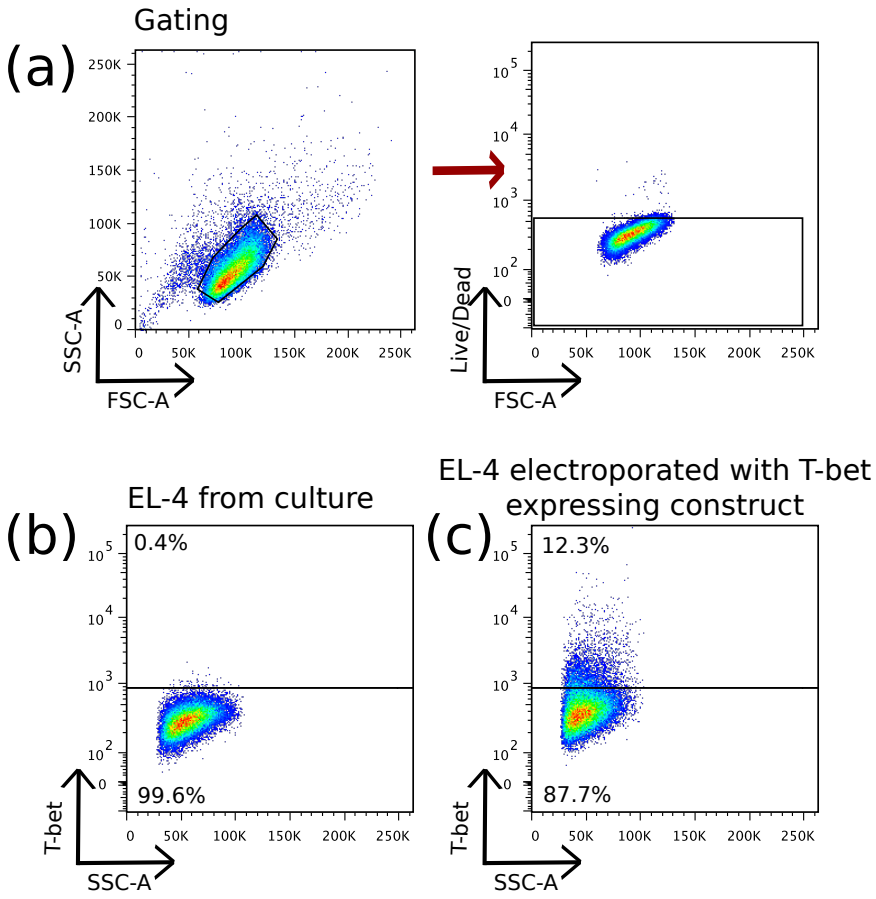

Figure 2.2: T-bet expression from T-bet expression construct. - EL-4 cells were electroporated with T-bet expression plasmid and then stained for T-bet. Cells were also stained with Live/Dead Aqua to allow for gating on live cells as seen in (a). (b) EL-4 cells from culture showed no T-bet expression. (c) Electroporated cells showed some T-bet expression.

2.8.3 Cell Transfection

EL-4 cells were triple transfected with 3 μg promoter construct, 6 μg expression construct and 0.3 μg renilla construct using the Nucleofector 4D system (Lonza). Transfection was as per manufacturers protocol except buffer used was manufacturers SG buffer and programme was CM-150 based on troubleshooting to improve viability and transfection efficiency. Each transfection was with 2×10^6 cells. Following electroporation, cells were rested for five minutes at 37°C in 500 μl RPMI before transfer to well of 24 well plate containing 1 ml pre-warmed DMEM. Cells were cultured for 20 to 30 hours before stimulation for 4.5 hrs with 50 ng/ml PMA and 1 μM Ionomycin. Cells were then harvested, washed once in PBS and luciferase assay was performed using Dual Luciferase Reporter Assay (Promega) as per manufacturers instructions. Luminescence was read on a Lumat LB9507 luminometer (Berthold Technologies).

2.9 Timecourse of IL-18 Signalling

Naïve cells from wild-type or T-bet$^{-/-}$ mice were harvested, FACS sorted, activated and cultured in Th1 skewing conditions for six days as described above. Cells were then harvested, washed once in PBS, once in X-Vivo 20 medium (Lonza, reference number BE04-448G) and then replated at 2×10^6 in 1 ml X-Vivo 20 medium. Cells were incubated for 90 minutes at 37°C and then 200 ng/ml of recombinant mouse IL-18 (R&D Systems, reference number B002-5) was added to one well at a time at 5 or 10 minute intervals over 40 minutes. Cells were then immediately harvested into ice cold PBS and centrifuged to wash. PBS was fully removed and cells were resuspended in 150 μl ice cold RIPA buffer (Sigma) which had been supplemented with 1 tablet of Complete Protease Inhibitor Cocktail per 10 ml RIPA and 1 $\times$ Phosphatase Inhibitor Cocktail 1 (Sigma, reference number P2850) and 1 $\times$ Phosphatase Inhibitor Cocktail 2 (Sigma, reference number P5726). Cells were incubated on ice for 30 minutes and then spun at 20,000g to pellet debris. Supernatant was snap frozen in liquid nitrogen and then stored at -80°C until use.

2.10 Western Blotting

Gels were precast 4% to 15% gradient gels (BioRad). Running buffer was 25 mM Tris-base, 200 mM Glycine, 0.1% SDS. Transfer buffer was 48 mM Tris-base, 39 mM Glycine (Fisher), 0.037% SDS (Fisher) and 20% Methanol. Tris-Buffered Saline supplemented with Tween 20 (TBS-T) was made from 10 $\times$ Tris-Buffered Saline stock (G-Biosciences) and supplemented with 0.1% Tween-20 (Sigma).

Oligonucleotide pulldown samples had been frozen in SDS loading buffer as detailed above and were denatured by heating from frozen at 95°C for 15 minutes before loading onto gel. Samples from IL-18 timecourse were left to thaw on ice and then 10 μl of 5 $\times$ SDS loading buffer, supplemented with 2-Mercaptoethanol, as described above, was added to 40 μl of sample and samples were heated at 95°C for five minutes. Gels were run in running buffer at 100 volts for approximately one hour before transfer to nitrocellulose membrane using transfer buffer. Transfer was conducted on ice at 60 volts for 1 hour 15 minutes. For oligonucleotide pulldown, samples were blocked in 5% milk in TBS-T and primary incubation was with 1:1000 αT-bet antibody (clone eBio4B10 from eBioscience) in 4% milk in TBS-T. For IL-18 timecourse samples, samples were blocked in 5% BSA in TBS-T and primary incubation was with 1:1000 αpp38 antibody (clone 12F8 from Cell Signalling Technology) in 5% BSA in TBS-T. For all experiments, blocking was for one hour at room temperature and primary incubation was at 4°C overnight. All incubation and wash steps were performed under gentle agitation. Blots were washed for 3 $\times$ 5 minutes in TBS-T at room temperature before addition of secondary antibody. Secondary antibody was 1:2500 anti mouse linked to Horse Radish Peroxidase (HRP) (GE Healthcare, reference number NA931) in 4% milk for oligonucleotide pulldown samples. Secondary antibody was 1:5000 αrabbit HRP (GE Healthcare, reference number NA934) in 5% BSA for IL-18 signalling timecourse samples. All secondary incubations were for one hour at room temperature. All blots were then washed three times with TBS-T before addition of Enhanced Chemiluminescent Substrate (PerkinElmer Inc, reference number NEL100001EA) as per manufacturers instructions and exposure to film. After development, timecourse blots were stripped with Restore stripping buffer (Thermo Scientific) for 25 minutes at room temperature under gently aggitation, washed three time in TBS-T, redeveloped to check for complete removal of signal and then reblocked as previ-

ously. Blots were then reprobed overnight at 4°C using 1:1000 α beta-actin antibody (polyclonal from Cell Signalling Technologies, reference number 4967S) in 5% BSA. Secondary incubation and redevelopment was as before. Blots were then restripped and reprobed for total p38 using 1:1000 α p38 antibody (polyclonal from Cell Signalling Technologies, reference number 9212P) in 5% BSA and using same procedure as already used for pp38 and beta-actin.

2.11 qPCR

Small bowel tissue sections from coeliac model were homogenised in 1 ml TRIsure (Bioline) using a TissueLyser (Qiagen) set to 25 Hz/s for five minutes. RNA was extracted by adding 200 μl chloroform to 1 ml TRIsure in tube, vortexing and incubating for 15 minutes at room temperature. Mixture was then centrifuged at 12,000g for 15 minutes at 4°C. After centrifugation, 500 μl Isopropanol and 0.5 μl GlycoBlue (Invitrogen) were added to the aqueous phase and this was transferred to a clean tube. The supernatant was vortexed briefly and then left at room temperature for 10 minutes. The sample was then centrifuged for eight minutes at 12,000g and 4°C to pellet the RNA. The RNA was washed twice in ice cold 75% ethanol and then left to air-dry fully. RNA was then resuspended in RNase free water and stored at -80°C until use. RNA concentration was measured with nanodrop. cDNA was made from RNA using Revert Aid First Strand cDNA Synthesis Kit (Thermo Scientific) according to manufacturers instructions. Reaction was performed using 1 μg of RNA and resultant cDNA was stored at -20°C until use. cDNA was diluted 1 in 2 prior to qPCR analysis. qPCR was performed using TaqMan probes in 384 well plate. Mixture was 5 μl Maxima Probe/ROX qPCR Master Mix 2 $\times$ (Thermo Scientific), 1 μl cDNA, 3.5 μl water and 0.5 μl gene specific probe mix (all from TaqMan). Probes incorporating the FAM dye were used for IL18R1 (Mm00515178_m1), IL18RAP (Mm00516053_m1), IFN-γ (Mm01168134_m1) and IL-17A (Mm00439619_m1). Probes incorporating the VIC dye were used for b-actin (4352341E). qPCR was performed on an ABI Prism 7900HT Sequence Detection System (Applied Biosystems) using a programme of 10 minutes at 95°C followed by 60 cycles of 95°C for 15 seconds and 60°C for 1 minute. Data was acquired and analysed on Sequence Detection Software version 2.4 (Applied Biosystems) Reactions were performed with one probe per tube based on preliminary data which

showed cross reaction between IL18RAP and b-Actin probes in multiplex reactions.

3

Trait-Associated SNPs in Binding Sites for the Transcription Factors T-bet and GATA3

3.1 Introduction

The recent development of large scale genomic analyses have demonstrated the importance of non-coding regions of the genome and some of the ways in which those regions can be analysed and annotated. In terms of understanding disease mechanisms, GWAS has provided many examples of DNA variation that fall outside of protein coding regions but that fall in other regulatory regions. Work to trace causal SNPs that alter disease risk and to understand the molecular mechanisms by which they exert their effect is ongoing. However, although large scale efforts such as the ENCODE[3] project have yielded vast amounts of information they have also highlighted the limitations of any such projects: we now understand the importance of context in genomic regulation and downstream effect. Possible connections between genetic variation, transcription factor binding, downstream biochemical effect and consequential clinical outcome, in terms of disease or increased risk of disease, have been postulated and explored. However, in many cases only partial explanations have been found or connections remain based on statistical association and lack detailed molecular mechanisms. To fully connect genetic variation with molecular mechanism and clinical outcome we need to search for the appropriate regulatory features in the

appropriate cell type and at the appropriate time. This combination is logistically daunting if not impossible in the broad-scope analyses of projects such as ENCODE. However by focusing a project, based on knowledge from other areas such as immunology, we can hope to narrow down the question sufficiently that appropriate cellular and kinetic contexts can be studied.

One mechanism by which a SNP can exert an effect is by altering binding of a transcription factor. To examine whether such altered binding can exert sufficient effect to alter the risk or course of a disease and to understand the mechanism by which it might do so, it makes sense to choose a transcription factor known to have a role in disease and immunity. Furthermore, we want to choose a cell type known to rely on the transcription factor for full function. The T helper cell lineages and their master regulators provide examples of this. Our laboratory has previously obtained ChIP-Seq data for the binding of T-bet in Th1 cells and GATA3 binding in Th1 and Th2 cells. T-bet is not expressed in Th2 cells. We decided to examine whether disease-associated SNPs from GWAS could be found in binding sites for T-bet in Th1 cells (hereafter just referred to as T-bet binding sites) or for GATA3 in Th1 or Th2 cells.

3.2 Generating a List of Binding Site SNPs

The NHGRI GWAS catalogue was accessed on 4$^{\text{th}}$ November 2011 and all SNPs listed in this catalogue (hereafter referred to as trait-associated SNPs) were downloaded. We used the snpMatrix package[300] to compile a list of all SNPs in high LD ($r^2 > 0.8$) with these trait-associated SNPs as given by the data in HapMap 3 for the CEU (Utah residents with Northern and Western European ancestry) population. Although not all GWAS in the NHGRI catalogue were performed in populations of European ancestry the prevalence of studies in this population made it an appropriate population for the first-pass analysis. We used HapMap data over whole sequence data because data from the 1000 Genome Project was not readily available at the time of establishing our *in silico* analysis. We had already moved on to examining further avenues of work when the 1000 Genome Project was published and chose to focus on functional analysis of the SNPs that we had already found rather than redoing our *in silico* analysis. We examined whether any of the trait-associated SNPs or the SNPs in high LD were in binding sites for T-bet or GATA3 as found by ChIP-Seq stud-

ies.[234] For the purposes of this work, strong LD is defined as $r^2 > 0.8$ unless stated otherwise. We defined strong LD as $r^2 > 0.8$ as a compromise. Using a higher r^2 value might have led us to miss potential hits, especially from earlier GWAS that were conducted across smaller numbers of SNPs. These earlier GWAS had a greater reliance on imputation through LD. By contrast, using a lower r^2 value might have generated too many false positives, increasing the number of hits to test in downstream analyses. Of note, $r^2 > 0.8$ was also used as a cut-off in one of the recent ENCODE publications.[107]

Our preliminary analysis generated 113 hits for T-bet (appendix table 7.1), 123 hits for GATA3 in Th1 cells (appendix table 7.2) and 89 hits for GATA3 in Th2 cells (appendix table 7.3). We considered a hit (hereafter referred to as a hit-SNP) to be any SNP in a binding site for the transcription factor under investigation that was in strong LD with a trait-associated SNP as taken from the NHGRI GWAS catalogue. We performed our initial systematic analysis on the autosomal chromosomes only.

3.3 Functional Genomic Annotation of Binding Site SNPs

3.3.1 Local Genomic Architecture - Histone Modifications and DNase Hypersensitivity

As discussed in the introduction, gene transcription is regulated at multiple levels and work on understanding those multiple levels is still ongoing. However, to start investigating whether any of these hits were likely to be biologically meaningful, we looked at whether any of the transcription factor binding sites that we found to contain a hit overlapped with any sites of histone modification. As previously discussed, H3K4me3 is a marker for transcriptionally permissive promoter regions, while H3K27me3 is a mark for regions that are transcriptionally repressed. Typically H3K4me1 is used as a mark for enhancer elements[313] but it also seen downstream of the TSS.[10] We downloaded published data on tag locations of the histone modifications H3K4me1, H3K4me3 and H3K27me3 from previously published work[10] and used this to call peaks of histone modification enrichment using the Model-based Analysis of ChIP-Seq (MACS) al-

gorithm. This algorithm uses a sliding window to calculate peak width and then searches for regions of peak-width that are enriched relative to a control sample or to local background.[302] We then found any T-bet or GATA3 sites that intersected with a peak of histone modification (table 3.1).

The number of transcription factor binding sites that overlapped with H3K27me3 was surprisingly small. However, we note that there were only 728 regions called for H3K27me3 in our analysis versus 28,074 for H3K4me3 and 45,913 for H3K4me1 (see table 3.1). Given that extra data was later released for H3K27me3 by the same group,[9] this would suggest that the small numbers result from technical reasons relating to the original H3K27me3 histone modification data rather then reflecting underlying biological mechanisms. Since the new data was not readily available in a standard .bed file format we were unable to incorporate the extra peaks into our analysis at this time. We reran the initial analysis to search for hit-SNPs that were just in the binding sites detailed in table 3.1. Unsurprisingly, we did not find any hits in the transcription factor sites that overlapped the H2K27me3 modification for either T-bet or GATA3 and we decided not to pursue this analysis any further. Hit-SNPs in transcription factor binding sites that overlapped the H3K4me1 or H3K4me3 modifications are described as Enhancer (for H3K4me1), Promoter (for H3K4me3) or Both (for both modifications present) in appendix table 7.7, column five.

We also examined whether any of our hits were located in sites of DNase hypersensitivity as this would highlight regions where the chromatin was looser and more accessible, possibly indicating that the region was being actively transcribed. To make the analysis as biologically relevant to our own data as possible, we downloaded peaks for DNase hypersensitivity in Th1 and Th2 cells which had been found by DNase-Seq as part of the ENCODE project.[3] The number of overlapping peaks is shown in table 3.1. Our analysis showed that over half (9,336/14,835) of the T-bet binding sites overlapped a region of DNase sensitivity: the numbers were also high for GATA3 binding in Th1 and Th2 cells (7,112/14,169 and 6,314/12,926 respectively). This is unsurprising given that the cell types were well matched and that loose chromatin is generally required for a transcription factor to bind and to mediate transcription. These results are also in agreement with previously published work on T-bet and GATA3 binding.[234] Hit-SNPs that were in transcription factor binding sites that overlapped a region of DNase hypersensitivity are annotated in appendix table 7.7, column

six).

	H3K4me1	H3K4me3	H3K27me3	DNase	Total Number of Peaks for Transcription Factor
T-bet	3,709	4,217	26	9,336	14,835
GATA3 in Th1 cells	4,623	4,889	18	7,112	14,169
GATA3 in Th2 cells	3,498	3661	20	6,314	12,926
Total Peaks Histone Modification or DNase	45,913	28,074	728	77,773(Th1) 92,923(Th2)	

Table 3.1: Number of transcription factor binding sites that overlapped sites of histone modification and DNase hypersensitivity. The number of transcription factor binding sites for each of T-bet, GATA3 in Th1 cells and GATA3 in Th2 cells that overlapped with regions of H3K4me1, H3K4me3, H3K27me3 and DNase hypersensitivity are shown. Last column of table shows total number of binding peaks for each transcription factor. Last row shows total number of DNase hypersensitive sites.

3.3.2 Proximity to Genes

We then used dbSNP to compile information on the location of our binding site SNPs relative to the nearest gene. As discussed in the introduction, enhancers may act on their nearest gene and/or they may act on genes many kbps away and even on different chromosomes. However, in general, a good correlation is seen between enhancer activity and the expression of genes up to 50kbp away.[314] Promoter elements, by contrast, are much closer to the gene on which they act. Therefore, we compiled three lists of genes. The first was a list of genes where a hit-SNP was in the transcribed region of the gene (either 5' UTR, 3' UTR, intron or exon), the second was a list of genes within 2kbp of each hit-SNP and the third was a list of genes within 50kbp of each hit-SNP. Results for each of the transcription factors are shown in appendix tables 7.4 (T-bet), 7.5 (GATA3 in Th1 cells) and 7.6 (GATA3 in Th2 cells). Because we were predominantly interested in immune mediators, the function of which we could assay *in vitro* and *in vivo*, the lists include only established protein coding genes and open reading frames as well as miRNAs and some other established noncoding RNAs. Pseudo, hypothetical or partial genes were not included and neither

were tRNAs. Validated and established protein coding genes with a LOC classification were included but all other LOC regions were not. The classification LOC is used by dbSNP and other NCBI databases when no other published symbol for a gene is available and prefixes the Gene ID as given by NCBI.

3.3.3 Proximity to other SNPs

GWAS examine a selection of SNPs across the genome and infer knowledge on SNPs in high LD with those SNPs that are studied. Thus a GWAS rarely guarantees to find causal SNPs but instead finds a collection of SNPs, all of which will be in high LD and one of which might be causal. We allowed for this in our preliminary analysis. However, because of the use of LD, we cannot conclude, without further investigation, that our hit-SNPs are any more likely to be causal than the trait-associated SNPs from the GWAS catalogue. The most important SNP might be another SNP in the same LD block as the hit-SNP and trait-associated SNP. SNPs may act through any of several different mechanisms including altering transcription factor binding but also including the alteration of a coding region. Many nonsynonymous SNPs do not alter protein function (see introduction). It is conceivable that a neutral nonsynonymous SNP could be in LD with a disadvantageous regulatory SNP. However, the presence of a nonsynonymous SNP in the same LD block as a hit-SNP would suggest that the hit-SNP was less likely to be causal. Therefore, we decided that we wanted to prioritise SNPs that were not in LD with nonsynonymous SNPs for the first rounds of our subsequent analyses. To do this, we generated a list of all the SNPs in strong LD with the GWAS SNPs and downloaded information on the functional class of all these SNPs from the SNP Function Portal.[315] Any SNPs in LD with SNPs causing nonsynonymous coding mutations are shown in appendix table 7.7 column eight.

3.3.4 Presence and Alteration of Motifs

We then examined whether any of the hit-SNPs were in or near a consensus sequence for the transcription factor under study. Although a PWM had been created for both T-bet and GATA3 from the original ChIP-Seq data, this does not mean that every ChIP-Seq peak contains a motif for reasons discussed in

the introduction. Our preliminary analysis had been performed by searching for SNPs in 200bp regions centred on the ChIP-Seq peak summit as called by the MACS peak finding algorithm. As such, our preliminary analysis was independent of underlying sequence. We had refrained from using PWMs for our initial search because of the caveats as to their use as discussed in the introduction. However, while the lack of a consensus sequence might not necessary rule out a binding site as not containing a relevant hit-SNP, the presence of a sequence might increase the likelihood of an already listed hit-SNP being genuine. We examined the presence of motifs using two methods.

In the first method, we obtained FASTA format sequences for the 20bp either side of each hit-SNP. We then chose two potential consensus sequences based on the PWM for T-bet[234] and the standard GATA motif for GATA3 and searched for these using the `dreg` tool[316] from the European Molecular Biology Open Software Suite (EMBOSS). For T-bet we searched for either TTxxCAC or CACAC and their complements. For GATA3, we searched for the standard GATA motif and its complement. In each case, we analysed the sequences as they had been obtained and then altered the nucleotide at the hit-SNP position to the opposite allele to check for the creation or deletion of a motif by the hit-SNP. A workflow for this analysis can be seen in figure 3.1. Hits are annotated in appendix table 7.7 column six.

Of note, the majority of our sequences did not contain a match to one of the motifs (fig. 3.2). There are two potential explanations for this. Firstly, we were not examining the 200bp of binding site sequence itself but a region 40bp long and centred on the hit-SNP which was somewhere within the binding site. A motif could have been present in the binding site sequence but outside the range of the sequence used to search for a consensus motif. Secondly, there may not have been a consensus sequence present even in the binding site sequence. The `dreg` tool searches for one or more given strings of DNA sequence similar to a standard text search but with the ability to input strings with degenerate bases. Although we picked TTXXCAC and CACAC based on conservation within the published PWM we may have lost important information in moving from a PWM to a search string. This highlights a further issue with the use of consensus sequences. Furthermore, not all sequences will contain a match to the motif in the first place.

We reran our analysis with all of the actual MACS-called binding site sequences

that contained a hit-SNP to test how many did contain a matching sequence, even if it was not near the hit-SNP. Our results showed that most but not all the binding site sequences contained a matching sequence (fig. 3.3).

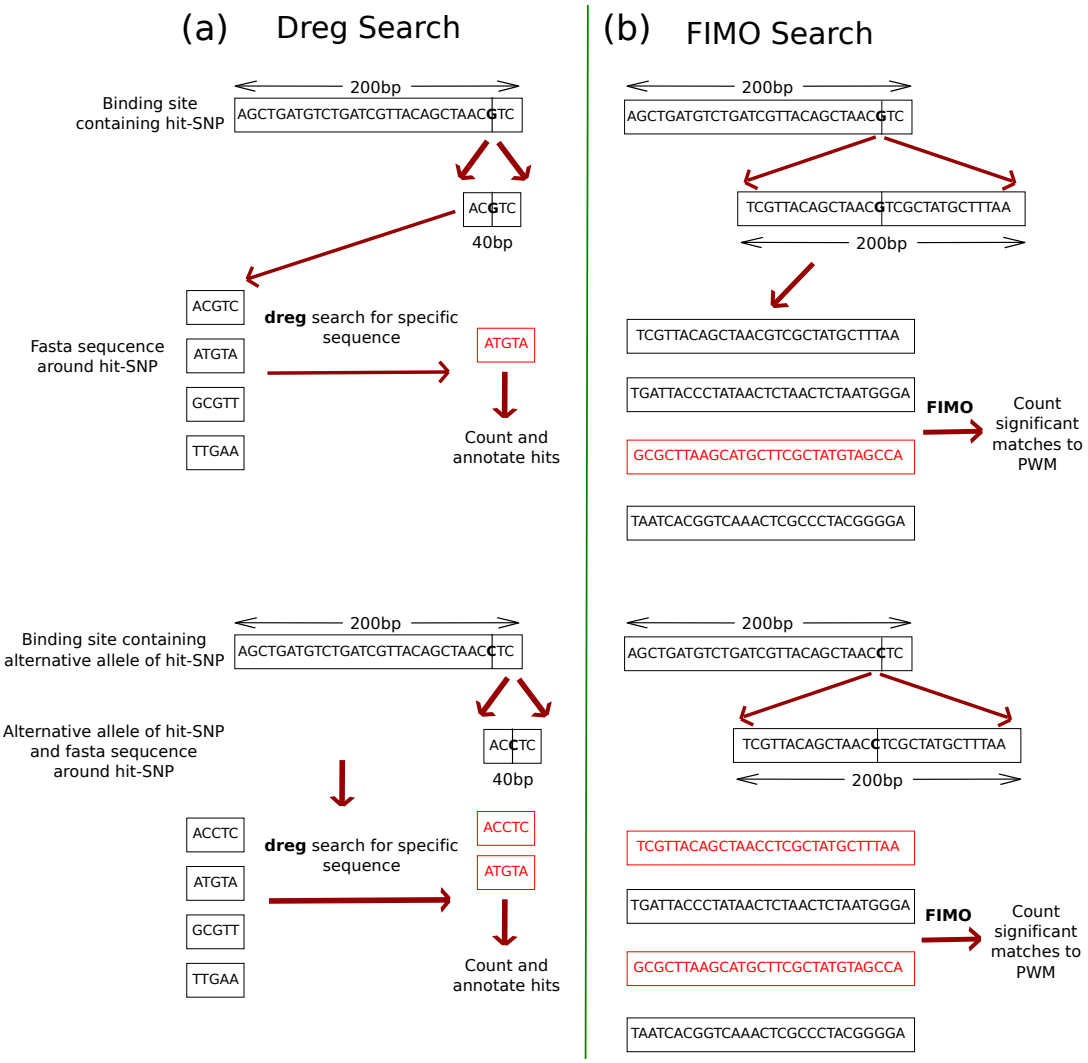

Figure 3.1: Work-flow for searching for consensus sequence or PWM around hit-SNPs. - Sequences around each hit-SNP were searched for a T-bet or GATA3 consensus sequence using the `dreg` tool (a) or for significant matches to PWM for T-bet or GATA3 PWM using FIMO (b). Workflow shows sequences tested relative to binding sites sequences.

In the second method, we obtained FASTA format sequences for the 100bp region either side of each of our hit-SNPs. Although these sequences were the same length as the binding site sequences, they did not correspond exactly to the binding site sequences that contained the hit-SNPs, because they were centred on the hit-SNP rather than the MACS peak summit (see fig. 3.1). We

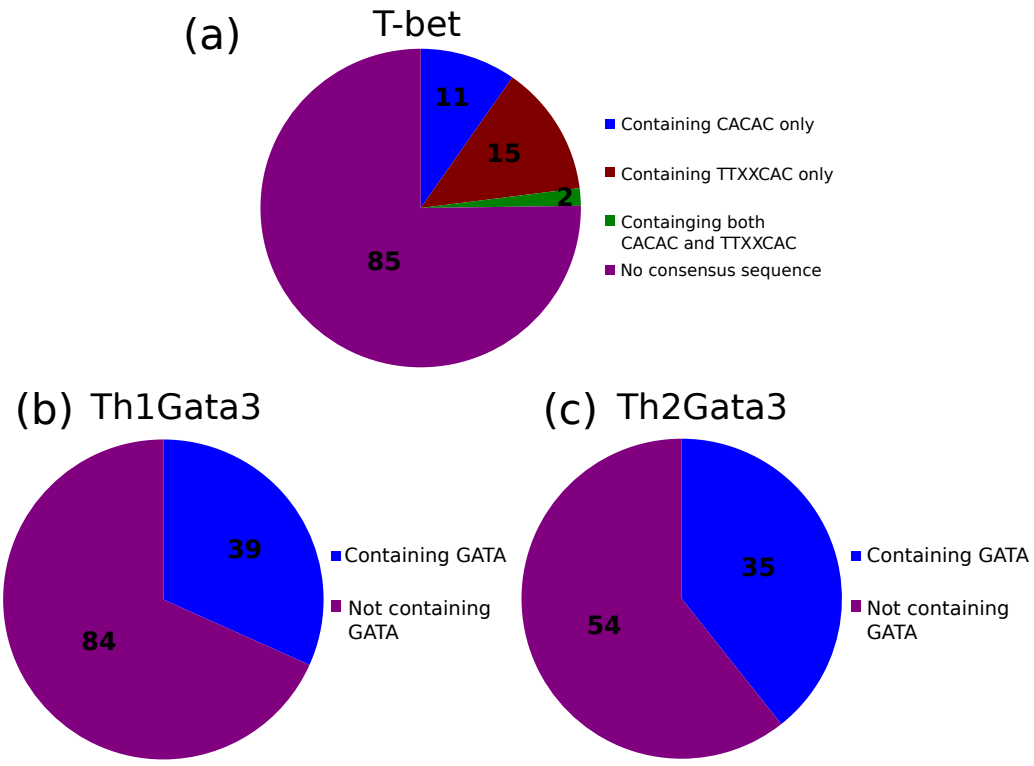

Figure 3.2: Majority of sequences around each hit-SNP did not contain a match to transcription factor consensus sequence. - Sequences around each hit-SNP were examined for consensus sequences taken from the PWM for T-bet and GATA3. Sequences around each hit-SNP in a T-bet binding site was examined for presence of TTXXCAC, CACAC or both (a). Sequences surrounding each hit-SNP in a GATA3 binding site in Th1 cells (b) or Th2 cells (c) were examined for GATA.

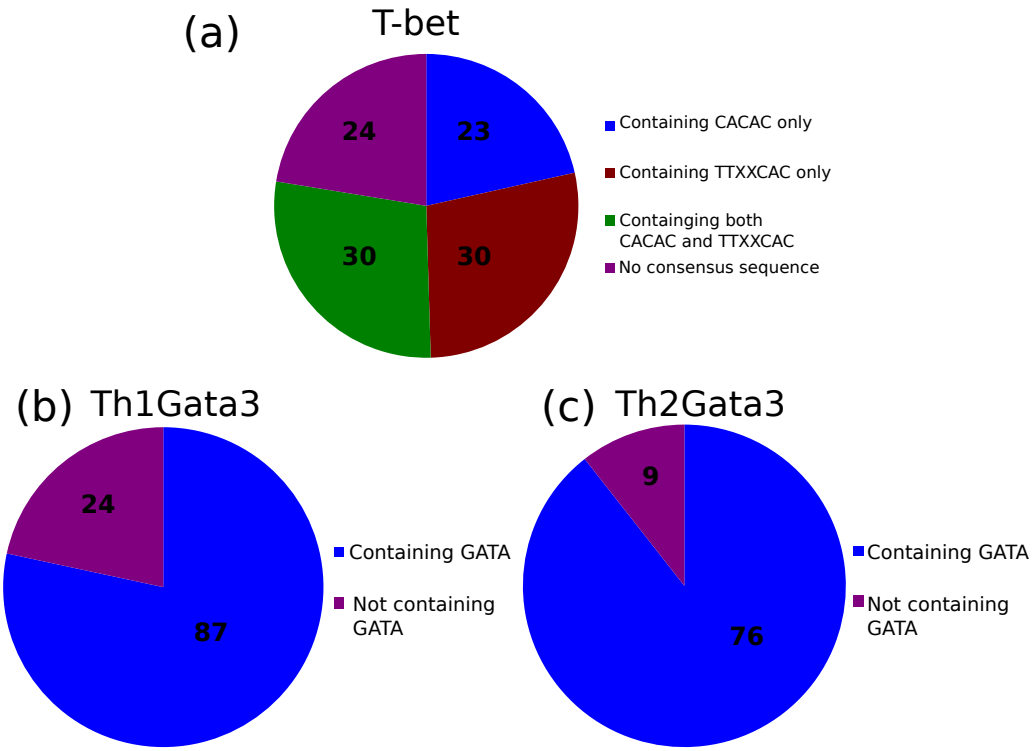

Figure 3.3: Most but not all binding site sequences containing a hit-SNP contain a consensus sequence as found by dreg. - Sequences of the binding sites containing each hit-SNP were examined for consensus sequences taken from the PWM for T-bet and GATA3. Sequences of T-bet binding sites were examined for presence of TTXXCAC, CACAC or both (a). Sequences of GATA3 binding sites in Th1 cells (b) or Th2 cells (c) were examined for GATA.

then used the PWMs generated for the T-bet and GATA3 binding sites from the ChIP-Seq data[234] and applied these to our sequences these using the Find Individual Motif Occurrences (FIMO) tool available through the EMBL Australia MEME Suite[305,306] (see materials and methods). This programme returned any sequences that matched the motif matrices with a p-value of less than 1×10^{-4} as calculated by a mixture of log likelihood ratios and dynamic programming as described by Grant *et al.*[305] The matches for the hit-SNP sequences are shown in appendix table 7.7 column six. Again, we noticed that not all of our sequences contained a significant match to the PWM. We tested how many of the binding site sequences containing a hit-SNP contained a match to the PWM. Again, we found that not all of the binding sites matched the PWM. Indeed, in this case, not even the majority of binding sites returned a significant match (fig. 3.4). However, there was enrichment of matches compared to a selection of sequences of the same length that were chosen randomly from the genome (fig. 3.4) validating the reliability of the analysis.

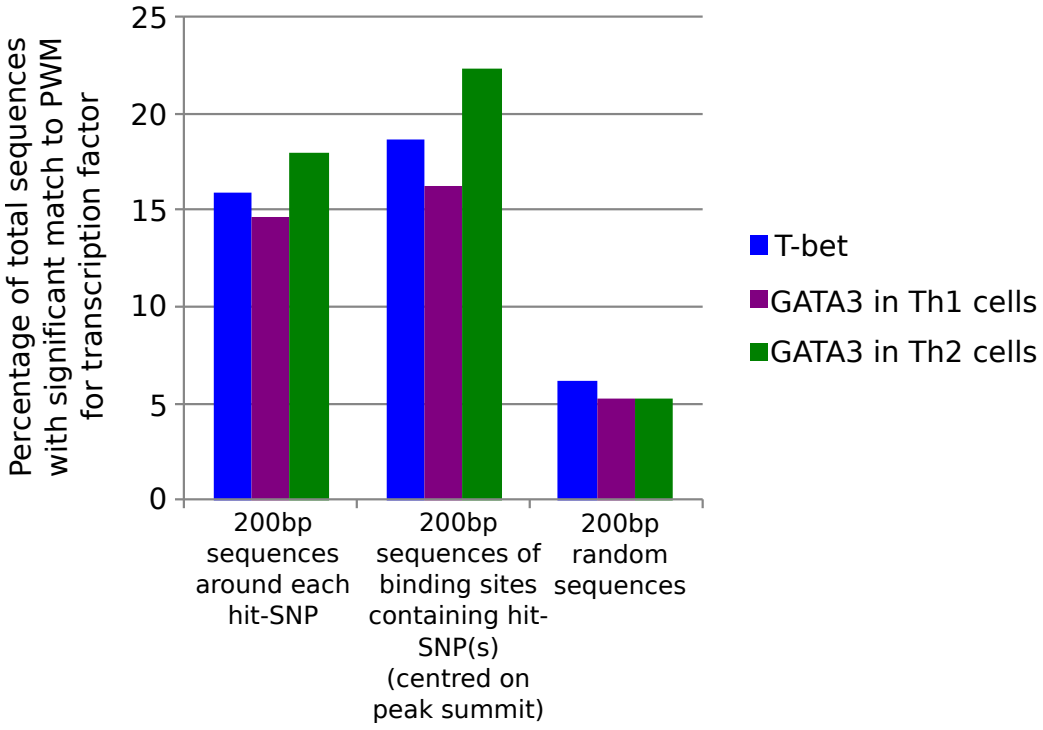

Figure 3.4: Percentages of sequences with significant matches to transcription factor PWM. - Sequences around hit-SNPs and sequences of actual binding sites containing hit-SNPs were examined for matches to the PWM for the transcription factor tested using FIMO. Enrichment for hits was tested by analysing with respect to random sequences.

3.4 Transcription Factor Co-binding

As we were performing the analyses, we noticed that nearly 30% of our hit-SNPs for T-bet were also in a binding site for GATA3 in Th1 cells. (fig. 3.5). This observation agrees with previous observations[158,234] that T-bet and GATA3 binding sites often overlap. The overlap in total binding sites on the autosomal chromosomes, independently of the presence or absence of a SNP, is shown in figure 3.6. The percentages of total binding sites that overlap with respect to total number of sites for each transcription factor are broadly similar to the percentages of SNPs found in multiple binding sites (fig. 3.7). The overlap between binding sites and hit-SNPs for GATA3 in Th1 and Th2 indicates that GATA3 binds to these locations in both cell types. Of more interest are those sites bound by GATA3 in Th1 but not Th2 and vice versa as these highlight regions of differential regulation involving GATA3 in the two different lineages. Kanhere *et al* have shown that T-bet can physically redistribute GATA3 in Th1 compared to Th2 cells.[234] Some of the overlapping regions between T-bet and GATA3 in Th1 cells could represent co-binding within the same cell. The overlap between T-bet binding sites and GATA3 binding sites in Th2 cells does not represent co-binding as T-bet is not expressed in Th2 cells. These sites may point towards differentially regulated loci within the two different cell lineages.

We wanted to examine whether any of the regions surrounding our hit-SNPs had, in common, the binding of any transcription factors other than T-bet and GATA3. To examine the possibility that other transcription factors might bind in the regions around our hit-SNPs and might be affected by the presence of the SNP, we searched for other consensus sequences in our sequences using the Multiple Em for Motif Elicitation (MEME) tool available through the MEME suite.[306] This analysis is not ideal as consensus motifs are usually built on thousands of sequences and we had a maximum of 123 sequences. However, we did find a Runx motif in the T-bet sequences and a weaker Runx motif in the sequences for GATA3 binding sites in Th2 cells (fig. 3.8). The motif was found in 21 (19%) of the sequences around a hit-SNP in a T-bet site and 42 (52%) of the sequences around a hit-SNP in a GATA3 site in Th2 cells. In the T-bet sequences, this motif was altered by the presence of SNP rs13333528. In the sequences for GATA3 binding in Th2 cells, this motif was altered by SNPs rs1775312, rs2058622, rs40452, rs7171233, rs1596017.

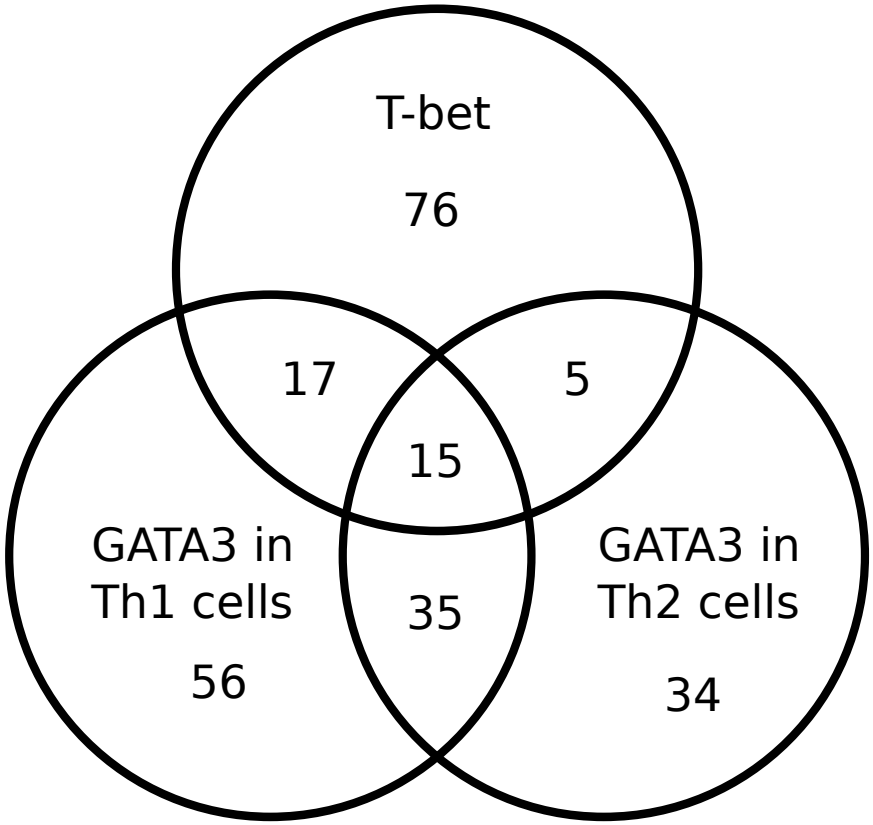

Figure 3.5: Hit-SNPs are found in multiple transcription factor binding sites. - Some of our hit-SNPs were found in binding sites for two or more of T-bet, GATA3 in Th1 cells and GATA3 in Th2 cells. Numbers of SNPs found in more than one transcription factor binding site are shown.

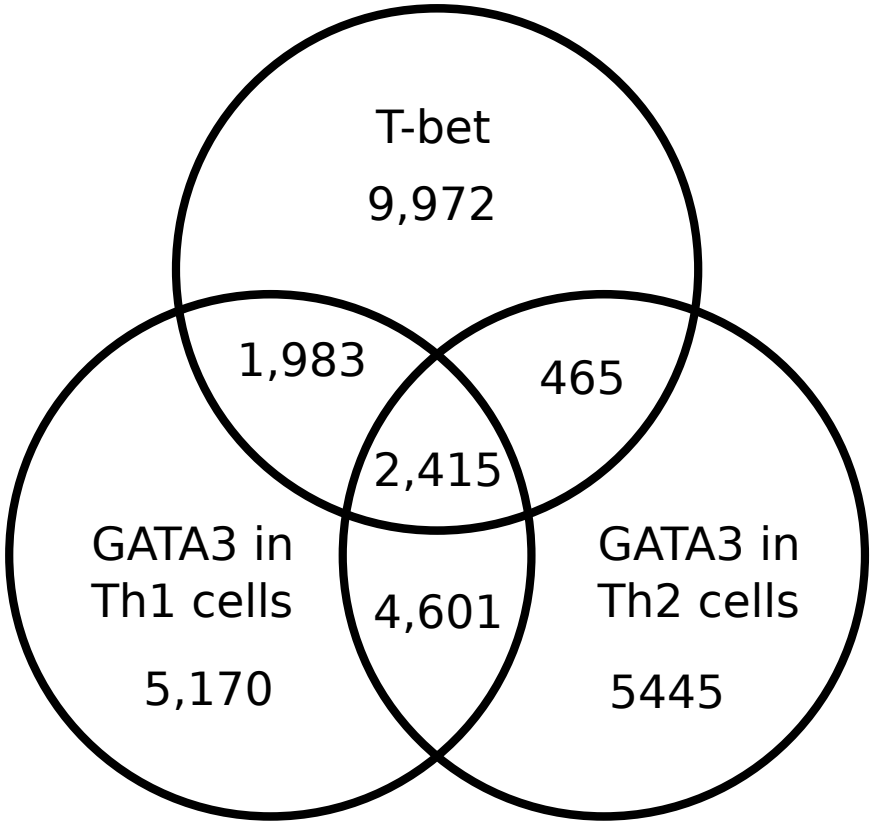

Figure 3.6: Transcription factor binding sites overlap. - Number of overlapping transcription factor binding sites for T-bet and for GATA3 in Th1 and Th2 cells out of total number of binding sites on the autosomal chromosomes.

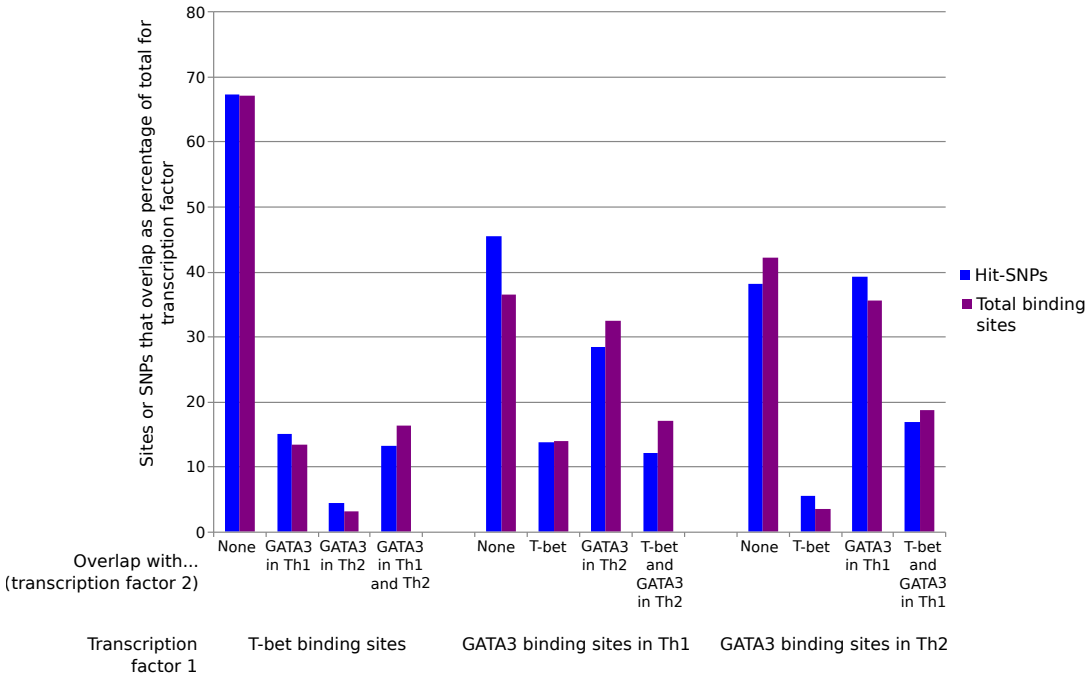

Figure 3.7: Overlap of hit-SNPs and transcription factor binding sites. - Overlap of total binding sites (purple) or hit-SNPs (blue) for each transcription factor condition (transcription factor 1) with each of the remaining transcription factor conditions (transcription factor 2). Percentages are out of total number of binding sites or hit-SNPs for transcription factor 1.

(a) T-bet

(b) Th2Gata3

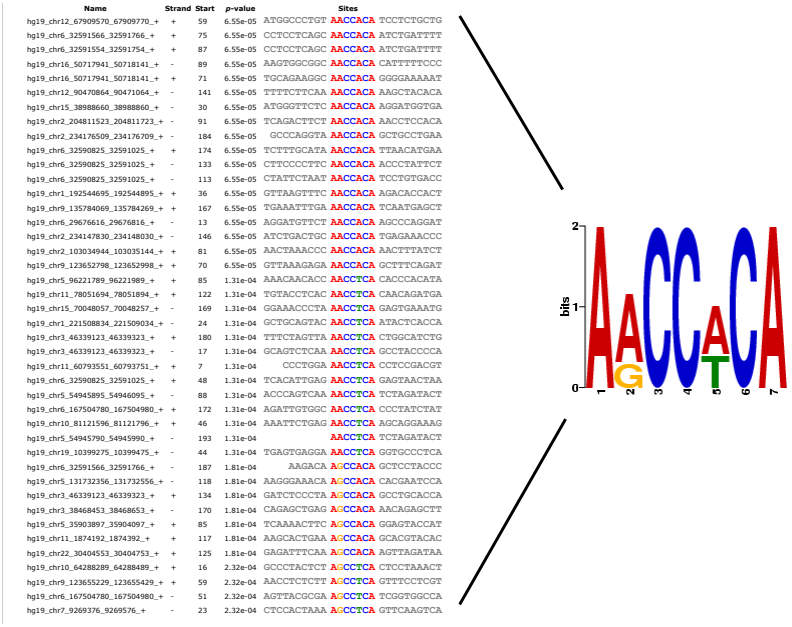

Figure 3.8: Runx motif found in some regions around the hit-SNPs for T-bet binding and GATA3 binding in Th2 cells. - The 200bp sequences around our hit-SNPs were analysed for motifs for other transcription factors and a Runx motif was found in the sequences around the hit-SNPs for T-bet (a) and GATA3 in Th2 cells (b). Sequences contributing to the motif are shown on the left and motif built from those sequences is shown on the right.

We also examined whether any other transcription factors might bind near our hit-SNPs using Transcription Factor Affinity Prediction (TRAP), a method which ranks transcription factor binding based on estimated binding affinities rather than searching for specific motifs.[307] For the sequences around the hit-SNPs for GATA3 binding in Th2 cells, the highest ranking motif was for GATA1, although the motif for GATA3 still reached significance above a control analysis using random sequences. GATA3 was also significant for the list of sequences around the hit-SNPs for GATA3 in Th1 cells. Because we were searching for DNA binding proteins against the JASPAR database, which does not contain information for T-bet, we were not expecting to see this motif returned as significant for the sequences surrounding the T-bet hit-SNPs. However, the T protein which is, like T-bet, a T-box protein and the homologue of *Xenopus* brachyury was significant. These results suggested that our hits were plausible. The significant binding events found are summarised in table 3.2.

T-bet	GATA3 in Th1	GATA3 in Th2
AP1	RUNX1	GATA1
RUNX1	SPI1	RUNX1
SPI1	GATA1	SPI1
IRF1	FEV	GATA3
FEV	ELF5	RORA
ELF5	IRF1	EVI1
T	T	FEV
NFE2L2	IRF2	ARID3A
RORA	EVI1	ELF5
IRF2	GATA3	IRF1

Table 3.2: Potential transcription factors binding around hit-SNPs. The sequences around our hit-SNPs for T-bet binding, GATA3 binding in Th1 cells and GATA3 binding in Th2 cells were analysed for other potential transcription factor binding events using TRAP. Transcription factors that scored more highly than the highest score assigned to a transcription factor binding event at a set of randomly generated control sequences are listed.

3.5 Immunological Annotation of Binding Site SNPs

We noticed that our hit-SNPs seemed to be enriched for binding site SNPs that were either associated with immune mediated diseases or had been generated because they were in high LD with SNPs that were associated with immune mediated diseases. To test this further, we divided the traits in the database into

three categories, traits which were definitely immune driven (such as coeliac disease and T1D), traits which were definitely non immune driven (such as mathematical ability) and traits which may or may not be immune driven (including various cancers). We performed this three- way classification as we wanted to analyse all SNPs but, at the same time, appreciated that not all traits can be split cleanly and definitely into immune and non-immune driven conditions. From this, we constructed a list of 'strictly' immune related trait-associated SNPs (hereafter referred to as strictly immune related SNPs) that were only associated with traits that were definitely considered immune mediated. We also generated a list of 'loosely' immune related trait-associated SNPs (hereafter referred to as loosely immune related SNPs) that were associated with at least one condition that might be immune mediated. We then counted how many SNPs from the list of strictly immune related SNPs had generated a hit-SNP. In other words, how many of the trait-associated SNPs from the strictly immune related SNP list were in high LD with a SNP in a binding site for T-bet of GATA3. We compared this number to the total number of strictly immune related SNPs in the entire GWAS catalogue. If there was no enrichment for immune mediated disease among those traits that had generated a hit-SNP then we would expect the proportion of strictly immune related SNPs that generated a hit-SNP to be similar to the proportion of strictly immune related SNPs in the entire NHGRI GWAS catalogue. However, we saw an enrichment of both strictly and loosely immune related SNPs that had generated a hit-SNP compared to the entire database (table 3.3). In the case of T-bet, for example, 40.1% of hit-SNPs were in LD with a strictly immune related SNP compared to only 13.9% in the entire database.

This data suggested that our results might have functional relevance given the important roles of both T-bet and GATA3 in the immune system. However, we might expect to see an enrichment even if the SNPs do not exert mechanistic effect through altering transcription factor binding for the following reason. There is an enrichment of T-bet and GATA3 binding sites at immune genes.[158] Furthermore, for those GWAS traits that are immune mediated we would expect an enrichment of SNPs associated with immune related genes. The enrichment we saw may simply reflect these two enrichments without being biologically relevant.

In addition, the above analysis was performed by counting trait-associated

	Total Number of Trait-Associated SNPs	Number Trait-Associated SNPs that are Strictly Immune Related	% Trait-Associated SNPs that are Strictly Immune Related	Number Trait-Associated SNPs that are Loosely Immune Related	% Trait-Associated SNPs that are Loosely Immune Related
T-bet	147	59	40.1%	68	46.2%
GATA3 in Th1 cells	153	64	41.8%	84	54.9%
GATA3 in Th2 cells	98	30	30.6%	41	41.8%
Total Database	5222	724	13.9%	1249	23.9%

Table 3.3: Enrichment of immune related trait-associated SNPs. Trait-associated SNPs from the GWAS catalogue that had generated (were in strong LD with) our hit-SNPs were classed as strictly, loosely or not immune related and proportions of immune related SNPs relative to all SNPs were calculated. Bottom line shows values for total database.

SNPs, which does not completely reflect the number of hit-SNPs, as our preliminary analysis found some hit-SNPs tagged by multiple trait-associated SNPs and some trait-associated SNPs tagging multiple binding site SNPs. This would have given us duplications in our counting given that it is the function of the binding site SNPs that we are interested in. This is why, for example, the total number of trait-associated SNPs for T-bet is 147 rather than the 113 hit-SNPs that we found for T-bet. It is also why we did not feel it relevant to subject the numbers given in table 3.3 to extensive statistical testing. Table 3.3 represents a quick analysis which was done to determine whether further analysis was worthwhile. Given that the data used in table 3.3 does not completely and accurately reflect the biological question being asked, we felt it would be misleading to start trying to determine statistical significance. It should be noted that table 3.3 suggests, rather than conclusively demonstrates, that there is an enrichment for immune related SNPs in T-bet and GATA3 binding sites. Enrichment within the data presented in table 3.3 could be analysed as immune related versus non immune related SNPs by chi-squared test but this would still fail to address some of the underlying biological issues that are discussed above. These issues are addressed in the following sections.

3.5.1 Randomisation of Binding Sites

We used two different methods to address some of the issues with our first analysis and to test the significance of the enrichments we saw. In the first method, we made lists of all SNPs that were strictly or loosely immune related and all the SNPs in strong LD with those SNPs and counted how many of these SNPs were in a binding site for T-bet. *In silico*, we then shifted each T-bet binding site in our list of binding sites along the chromosome by a random number of base pairs between 500bp and 5000bp either up or downstream of the actual binding site. We then examined how many of our immune related SNPs and their LD counterparts were in the newly shifted T-bet sites. By shifting between 500bp and 5000bp, we had created new sites that were associated with the same genes but not in the actual biological binding sites. If the enrichment in our initial T-bet hits was purely due to pre-enrichments based on proximity to immune related genes, we would expect to see a similar enrichment for immune related SNPs in the shifted T-bet sites. We repeated the binding site shifting and re-analysis through 10,000 random re-assignments of binding site position to test for significance of the enrichment. We then repeated for GATA3 in Th1 and Th2 cells.

This analysis did suggest that the enrichment for immune related SNPs in T-bet binding sites is biologically relevant and suggested strongly that the enrichment was also relevant for GATA3 in Th1 cells (table 3.4). Significant enrichments were not seen for GATA3 in Th2 cells ($p > 0.05$ for Th2Gata3 binding site randomisations in both cases.)

There are two potential issues with this analysis. The first is potential bias due to unequal coverage of traits in the GWAS catalogue. There is bias in the traits that have been studied and this is reflected in a bias in traits listed as associated with SNPs in the catalogue. In particular common traits that gave many hits in preliminary GWAS such as Crohn's disease have been further studied with larger cohorts and better statistical power than other traits, which have received less attention. In addition, some traits that are harder to study functionally, such as various psychological conditions, have been extensively subject to GWAS. These could be overrepresented in the GWAS catalogue as a result. Furthermore, in some conditions it is possible that the population in the initial GWAS was heterogeneous for whether their condition was immune mediated. Some cancers, for example, can give the same clinical result but may be driven

by different biological processes (inflammation, environmental exposure etc). This level of detail may not always be reflected in the original GWAS paper or our subsequent meta-analysis through the NHGRI GWAS catalogue. Because our *in silico* experiments were broad, we may have not captured finer details from individual papers.

3.5.2 Permutation of SNPs

As another way to assess the biological relevance of our analyses, we reassigned the immune status of our SNPs but used the real transcription factor binding sites. We counted the number of trait-associated SNPs from the NHGRI GWAS catalogue that we had called strictly or loosely immune related. We then chose this number of SNPs at random from the entire list of SNPs in the GWAS catalogue. The number of these permuted SNPs, or SNPs in high LD with these permuted SNPs, that were within the actual binding sites for each of our transcription factors was then counted and averaged over 200 permutations. Because this analysis was more computationally intensive than the analysis of shifted binding sites we were not able to test the significance of our results by standard permutation testing. Instead, we checked that the number of hits obtained over 200 permutations was approximately normally distributed and then checked for significance of our enrichment using a standard z-test over the 200 permutations. The results are shown in table 3.4. This analysis found significant enrichment of strictly immune related SNPs for all three transcription factor conditions ($p < 0.05$ in all cases.) Graphs showing the normal distribution are shown in figure 3.9.

Testing for significance by permuting the immune status of the SNPs is, we could argue, more statistically robust than shifting the transcription factor binding sites. The lists of permuted SNPs are chosen independently of their actual immune status. By contrast, the shifted binding sites are chosen to be a certain distance away from their actual location. The distance by which each site is shifted is random each time but within the constraints that it must be within 500 to 5000bp of the original location and on the same chromosome - the new location is not chosen at random from the entire genome. Furthermore, as a result of this, we noticed that we would sometimes pick up the same LD block and same disease-associated SNP for the shifted transcription factor binding

site as for the actual transcription factor binding site. However, the purpose of shifting the binding sites was to test whether the enrichment was specific to the binding sites and not to the genomic locations in which they were found. In answering this important biological question, the shifting analysis was more relevant.

	Actual Number Immune Related SNPs in Transcription Factor Binding Sites	Average Number SNPs in Shifted Transcription Factor Binding Sites	Average Number Permuted Immune Related SNPs in Transcription Factor Binding Sites
T-bet (Strict)	42	31 (p = 0.0359) (s.d = 5.6)	25 (p = 0.0006) (s.d = 5.3)
T-bet(Loose)	50	39 (p = 0.0543) (s.d = 6.3)	41 (p = 0.0795) (s.d = 6.2)
GATA3 in Th1 cells (Strict)	48	27 (p = 0.0002) (s.d = 5.2)	28 (p = 0.0005) (s.d = 6.2)
GATA3 in Th1 cells (Loose)	62	34 (p = 0.0001) (s.d = 5.9)	47 (p = 0.0285) (s.d = 7.9)
GATA3 in Th2 cells (Strict)	28	24 (p = 0.2026) (s.d = 4.9)	18 (p = 0.0128) (s.d = 4.7)
GATA3 in Th2 cells (Loose)	37	30 (p = 0.1284) (s.d = 5.6)	30 (p = 0.1382) (s.d = 6.5)

Table 3.4: Hit-SNPs are enriched for immune related SNPs. Enrichment of immune related SNPs was tested by randomly shifting transcription factor binding sites and by permuting immune status of GWAS SNPs. Actual number of strictly and loosely immune related SNPs in transcription factor binding sites is shown in column two. Column three shows mean and standard deviation of strictly and loosely immune related SNPs in shifted transcription factor sites. P-value was obtained over 10,000 randomisations. Column four shows mean and standard deviation of strictly and loosely immune related SNPs in transcription factor sites over 200 permutations of immune status reassignment. P-value obtained by z-test over 200 permutations.

3.5.3 Application to other Transcription Factors

To examine whether enrichment for immune related SNPs could be seen in other transcription factors, we obtained data on the binding sites of NF-κB, STAT4 and Oestrogen Receptor alpha (ER-α) and subjected these to our analyses. NF-κB was chosen as our positive control, a transcription factor known to be involved in immune responses. ER-α was chosen to be our negative control as this is not thought to be heavily involved with immune response. STAT4 is a

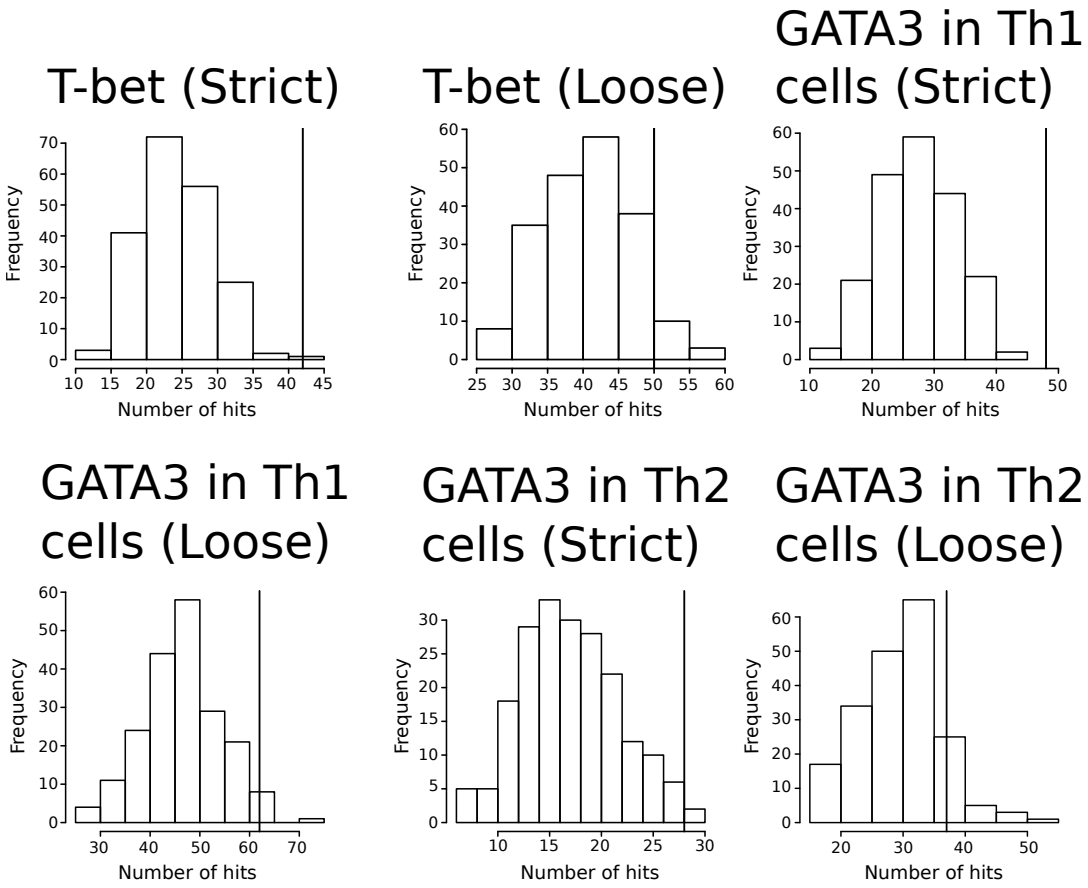

Figure 3.9: Number of permuted hit-SNPs followed approximately normal distribution. - Number of hits recorded for each re-assignment of immune related SNPs was plotted as a histogram for each transcription factor and for choosing a number of SNPs corresponding to the number of strictly immune related SNPs (strict) and number of loosely immune related SNPs (loose). Vertical line indicates actual number of strictly or loosely immune- related hit-SNPs for each transcription factor.

very general transcription factor which has functions in the immune system as discussed in the introduction. The results of our testing on these transcription factors are summarised in table 3.5.

	Actual Number Immune Related SNPs in Transcription Factor Binding Site	Average Number SNPs in Shifted Transcription Factor Binding Sites	Average Number Permuted Immune Related SNPs in Transcription Factor Binding Sites
NF-κB (Strict)	429	293 ($p < 0.0001$) (s.d = 16.7)	247 ($p < 0.0001$) (s.d =27.6)
NF-κB(Loose)	573	407 ($p < 0.0001$) (s.d =20.0)	425 ($p < 0.0001$) (s.d =30.5)
STAT4 (Strict)	67	52 ($p = 0.0172$) (s.d = 7.1)	32 ($p < 0.0001$) (s.d = 6.5)
STAT4 (Loose)	81	63 ($p = 0.0187$) (s.d = 8.0)	55 ($p = 0.0003$) (s.d = 7.5)
ER-α (Strict)	14	9 ($p = 0.0965$) (s.d =3.2)	11 ($p = 0.2435$) (s.d = 3.7)
ER-α (Loose)	20	14 ($p = 0.0764$) (s.d = 3.8)	21 ($p = 0.5630$) (s.d = 4.5)

Table 3.5: Summary of immune related SNPs in binding sites for other transcription factors. Summary of immune related SNPs in binding sites for NF-κB, STAT4 and ER-α. Data obtained and presented as for T-bet and GATA3.

Our results showed a strong enrichment for our positive control NF-κB. By contrast our negative control ERa showed no significant enrichment supporting the methods we had used and the classification we had made. Of interest, STAT4 also showed significant enrichment of actual immune related hit-SNPs in actual sites versus permuted SNPs or SNPs in shifted immune sites.

3.6 Algorithm Comparison

Although most of our analysis was done using T-bet and GATA3 binding site peaks as called by the MACS algorithm, we also had data on the binding site peaks for the same ChIP-Seq experiments as called by the Site Identification from Short Sequence Reads (SISSRs) algorithm[317] which calls peaks based on the directionality of aligned reads. In general, most peaks called by SISSRs will also be called by MACS but only around 50% of peaks called by MACS will be called by SISSRs.[22] Therefore, although MACS is considered by many as the

gold standard, we ran a comparison. Peaks called by both algorithms are highly likely to represent real transcription factor binding events which would give us increased confidence in any hit-SNPs in these binding sites. However, such a comparison would also allow us to examine whether either of the algorithms was better for calling binding sites that might contain hit-SNPs. SISSRs tends to call fewer peaks. However, if these are the peaks of strong binding that are more likely affected by a single base pair change, then using this algorithm, for our purposes, would save effort in downstream analyses. However, if SISSRs missed peaks containing hit-SNPs that were later validated then this would confirm the usefulness of MACS.

As can be seen from figure 3.10, the overlap between hit-SNPs when the binding sites are called by SISSRs and hit-SNPs when the binding sites are called by MACS is not very strong. This is in agreement with the work by Wilbanks *et al*[22] discussed above. Without an independent way of verifying which peaks represent true binding sites we cannot be certain as to whether the SISSRs algorithm is not sensitive enough and misses peaks or whether the MACS algorithm is not specific enough and returns many false positives.

3.7 Specific Analyses of the X Chromosome

We conducted most of our analysis on the autosomal chromosomes. However, binding regions on the X chromosome were also found in our ChIP-Seq data for all three datasets, T-bet, GATA3 in Th1 cells and GATA3 in Th2 cells. The chromosome harbours important genes including the Treg master regulator FoxP3 and mutations on this chromosome are known to cause diseases such as haemophilia and, in relation to the FoxP3 gene in particular, Immunodysregulation Polyendocrinopathy and Enteropathy X-linked (IPEX). Furthermore, GWAS hits have been reported on the X chromosome associated with a wide variety of traits including height, smoking behaviour, coronary heart disease and prostrate cancer. We repeated our basic analysis to look for SNPs that were in a binding site and either in the GWAS catalogue or in high LD with a SNP in the catalogue but found no hits. Therefore, we did not pursue analysis on this chromosome any further.

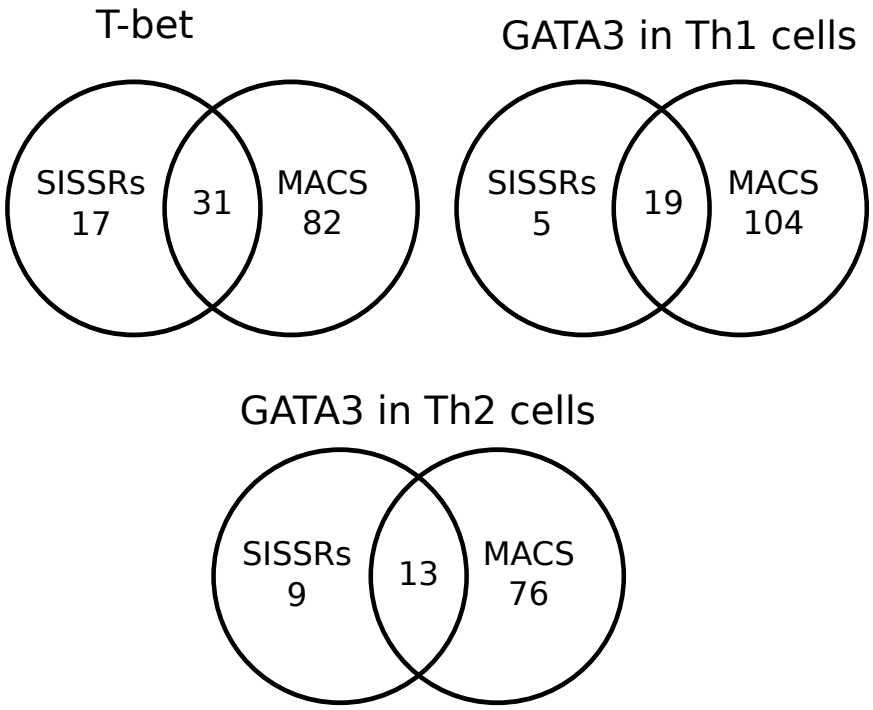

Figure 3.10: Overlap between SNPs in binding sites called by SISSRs and SNPs in binding sites called by MACS. - Binding site peaks for T-bet and GATA3 in Th1 and Th2 cells as called by the SISSRs algorithm were analysed for hit SNPs and the overlap of hit-SNPs called by the two algorithms is shown.

3.8 Conclusion

We have found multiple disease or trait-associated SNPs that are either in a binding site for T-bet or GATA3 or in high LD with a SNP in a binding site. Many of these hit-SNPs are found in regions marked as functionally relevant by genomic features such as histone modification or DNase hypersensitivity. Some of our hit-SNPs are near or in a motif for T-bet or GATA3 and some are in or near motifs for other transcription factors. Furthermore, some of our SNPs are in binding sites for both T-bet and GATA3.

It is unlikely that all of our hit-SNPs alter transcription factor binding and exert an effect on disease risk. However, we have some confidence in our results from the observation that our hit-SNPs were generated by traits that were enriched for immune mediated conditions. Hit-SNPs for T-bet were significantly enriched for strictly immune related SNPs as tested both by randomly shifting the T-bet binding sites and by randomly permuting the SNPs tested ($p < 0.05$ in both cases). Hit-SNPs for GATA3 in Th1 cells were significantly enriched for both strictly and loosely immune related SNPs. Of note, there was a far weaker enrichment of immune related SNPs in GATA3 binding sites in Th2 cells. This may reflect the biological differences in specificity of the two transcription factors. While T-bet is immune restricted, GATA3 has a broader role in biological processes.

4

In Vitro Testing of SNPs Found by *In Silico* Analysis

4.1 Introduction

Our *in silico* analysis revealed multiple SNPs that were in a binding site for T-bet or GATA3 or both and either associated with or in high LD with a SNP associated with a trait or disease in the NHGRI GWAS catalogue. However, although our genomic annotation and statistical testing of these hit-SNPs are encouraging, we have done no more than make associations at this time. To make causal links between our hit-SNPs and altered transcription factor binding and downstream mechanistic effect, we need to assess the function of our SNPs both *in vitro* and *in vivo*.

4.2 Some SNPs Show Altered Binding

4.2.1 Oligonucleotide Pulldown

Having found many hit-SNPs for both T-bet and GATA3 *in silico*, we started to investigate whether any of these SNPs altered transcription factor binding using oligonucleotide pulldown assays. We started by testing the effects of SNPs found in T-bet binding sites because a high affinity antibody is available for T-bet.[229] Biotinylated double stranded DNA oligonucleotides (hereafter referred

to just as oligonucleotides) containing each allele of the SNP of interest were bound to streptavidin beads. These beads were then incubated with lysate from YT cells which constitutively express T-bet. Bound protein was eluted from the beads and probed for T-bet by Western blot. We also designed a positive control (T-bet +), which contained a T-bet consensus sequence as recently determined from ChIP-Seq data on T-bet binding in Th1 cells.[234] We designed a negative control (T-bet -) in which this T-bet consensus sequence was mutated at two key residues. Using this technique, we tested hit-SNPs near or in the genes for IL18R1 (rs1465321), CCR1 (rs3181080) and RGS1 (rs2984920). We quantified the Western blots and found differential binding between the two alleles of rs1465321 (fig 4.1 (a) and (b)), with the G allele consistently binding T-bet more strongly than the A allele. We tested this SNP further by using lysate from bead enriched human $CD4^+$ cells that had been cultured *in vitro* in Th1 conditions for seven days. This was because transcription factor binding often depends on the presence of other transcription factors and on post-translational modification of the transcription factor such as phosphorylation. It is possible that the YT cell line, of NK lineage, does not represent Th1 cells in one of these aspects. Again, we saw stronger T-bet binding to the G allele of rs1465321 than to the A allele. By contrast, we found very little T-bet binding at the other hit-SNPs rs3181080 (fig. 4.1 (c)) and rs2984920 (fig. 4.1 (d)). Of note, figure 4.1 shows bars for individual experiments as the inter-assay variability was high between replicates. As such, we were more interested in consistency of differential binding (seen for rs1465321 but not rs3181080 or rs2984920) than in numbers for absolute band intensity. However, for rs1465321, if we try and control for inter-assay variability by normalising values for band intensity of the G allele to (band intensity of A allele =1), then a significant difference is seen across the combined experiments ($p = 0.049$, paired samples t-test).

4.2.2 OligoFlow

From the *in silico* analysis there were many SNPs that we wanted to test for differences in transcription factor binding, most of which were likely to be subtle. However, a conventional pulldown assay does not directly measure transcription factor binding to the beads but presence of the transcription factor in the final sample. This can introduce experimental error in the procedure. If the beads are not thoroughly washed or they are washed to different levels in different

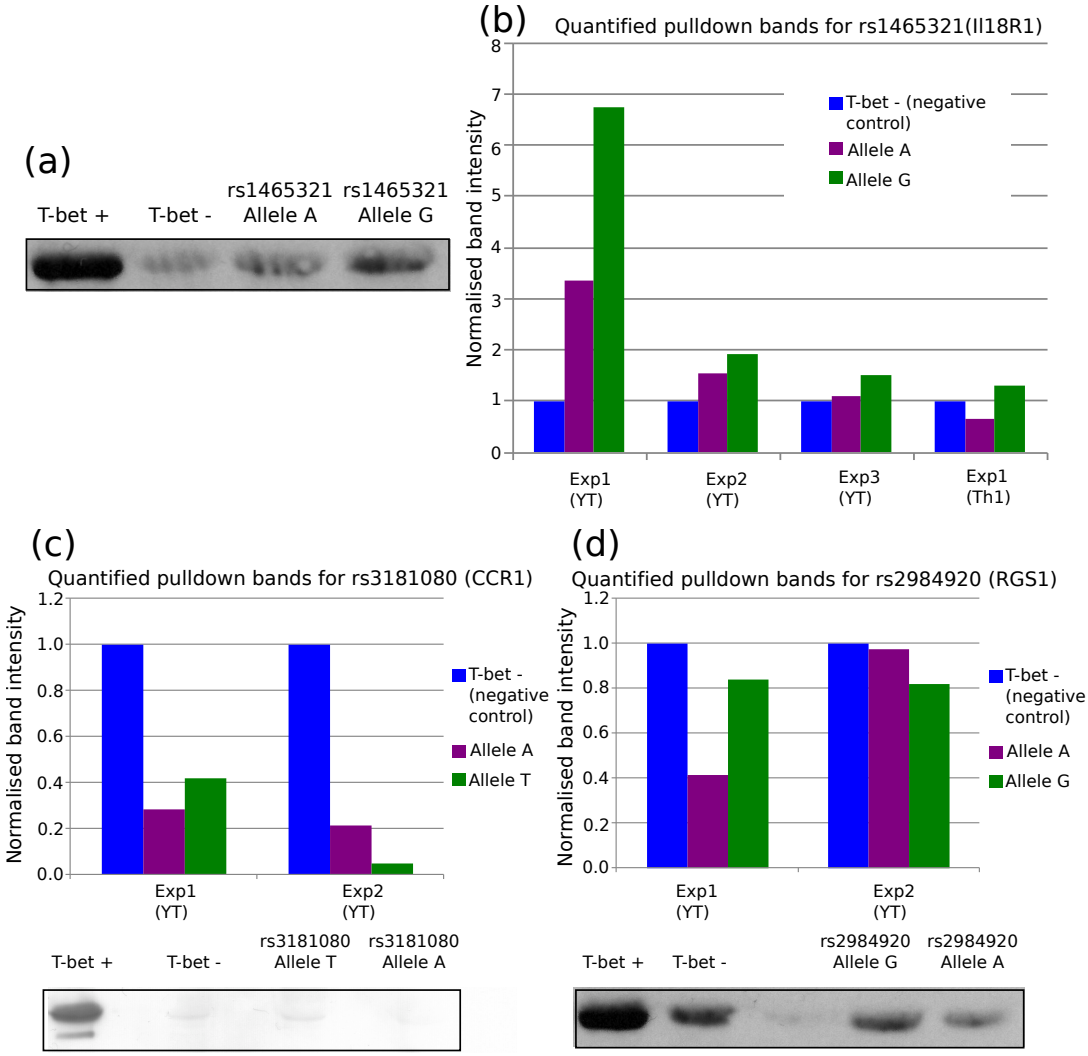

Figure 4.1: Consistent differential binding at rs1465321 but not rs3181080 or rs2984920. - Hit-SNPs for T-bet were tested for differential T-bet binding by oligonucleotide pulldown assay. (a) Example Western blot showing T-bet eluted from streptavidin beads prepared with positive control (T-bet +) or negative control (T-bet -) oligonucleotides, or oligonucleotides for the DNA region around rs1465321. (b) Quantification of Western blots from three separate pulldown experiments with rs1465321 and cell lysate from YT cells and one pulldown with rs1465321 and lysate from $CD4^+$ cells cultured in Th1 conditions *in vitro*. Results are normalised to (T-bet -) = 1 for each experiment. (c) Example Western blot and quantification of Western blots from two separate pulldown experiments with the DNA region around rs3181080. Data analysed as for (b). (d) Example Western blot and quantification of Western blots from two separate pulldown experiments with the DNA region around rs2984920. Data analysed as for (b).

samples or if any of the beads are lost during the wash steps then this will skew the results. This is especially problematic given that, unlike most other Western blotting procedures, there is no accepted internal control for which we can reprobe on the Western blot. In addition, a conventional pulldown assay is time consuming because it requires two steps, the assay itself and then the Western blot to analyse the results. This makes it difficult to scale up the assay to analyse SNPs in a medium to high throughput way. The issue of potential experimental variability is addressed in part by running the experiment multiple times but, in order to screen our hit-SNPs with better accuracy, we decided to develop a flow-based method for analysing pulldown samples. Flow cytometric analysis is a one step process that could easily be added onto the end of the pulldown assay, yielding data in approximately one hour versus the two day process of a Western blot. We reasoned that we could perform the pulldown assay as standard but, after the lysate incubation step, we could add a fluorescently labelled antibody for the transcription factor of interest instead of washing the beads, eluting protein and Western blotting. We could then measure the relative Median Fluorescence Intensity (MFI) as a readout of strength of transcription factor binding in each individual sample. An outline of this approach in comparison to a standard pulldown followed by Western blotting is shown in figure 4.2. A broadly similar idea has been used to study interactions between RNA binding proteins and their RNA target.[318] However, this previous work used GFP labelled protein which had to be synthesised from an appropriately constructed plasmid and then purified. In our case, we wanted to test whether we could achieve a signal using cell lysate containing the transcription factor of interest in its native form and an appropriate antibody. We reasoned that, if we were able to do this, then this would broaden the applicability of the technique. The technique could be applied easily to test differential binding of a range of transcription factors. Furthermore, we wanted to examine whether the assay was sensitive enough to detect differences in binding as a result of single base pair changes in the oligonucleotide used.

To test this technique of running a pulldown assay and then analysing by flow cytometry (hereafter referred to as OligoFlow), we first used oligonucleotides for just the positive and negative controls for T-bet binding (T-bet + and T-bet-). We found that we needed to use streptavidin coated polystyrene beads rather than the standard agarose beads that are used when Western blotting as the agarose beads could not withstand the pressure of a flow cytometer fluidics

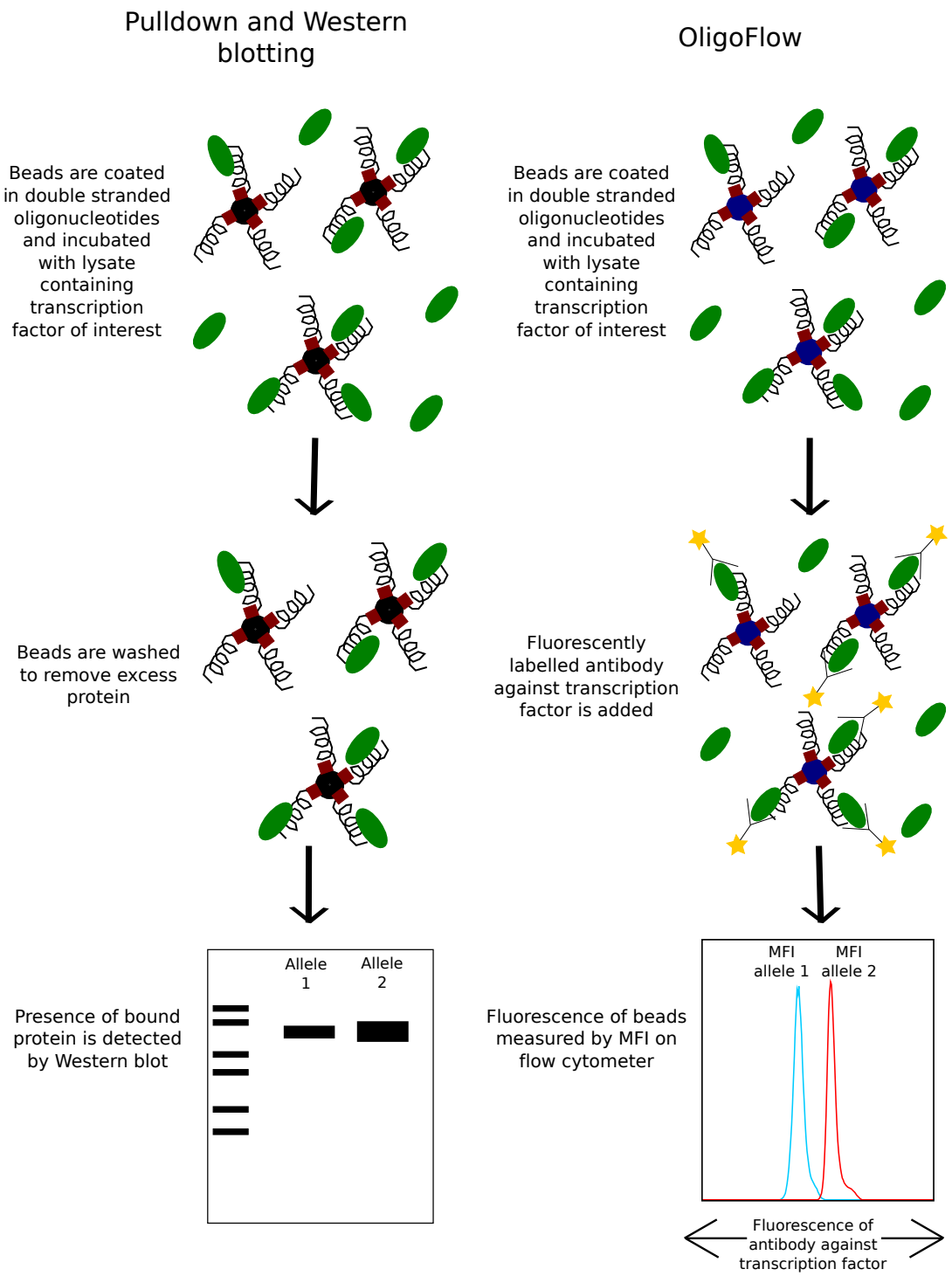

Figure 4.2: Overview of OligoFlow method. - Steps required for assessing differential transcription factor binding to DNA by OligoFlow in comparison to pull-down assay followed by Western blotting.

system. We were unsure as to whether the various available fluorochromes would be able to tolerate the buffer that is used for incubating the pulldown beads with cell lysate (the oligo buffer). Therefore, we performed a side by side comparison. We attached our positive or negative control oligonucleotides to beads and incubated with lysate. We then either added the antibody directly, incubated and acquired or washed the beads and resuspended them in PBS before incubation with the antibody. At the same time, we acquired FacsComp beads that had been added either to PBS or to an aliquot of the cell lysate and oligo buffer prior to incubation with the antibody. For our first analyses, we chose the Alexa647 fluorochrome as it is relatively stable and bright.

Incubating the FacsComp beads in the buffer used for oligonucleotide pulldown did decrease the signal intensity from our αT-bet Alexa647 antibody compared to FacsComp beads in PBS (fig. 4.3 (a) and (b)). However, it did not completely abrogate the signal and two peaks for positive and negative beads could still be seen (fig. 4.3 (b)) . When we compared the positive and negative control beads incubated in oligo buffer, we could see a marked increase in T-bet Alexa647 fluorescence on the beads made with positive control oligonucleotide (T-bet + beads) versus beads made with the negative control (T-bet - beads) (fig. 4.3 (e)) and versus background of polystyrene beads only (fig. 4.3 (d)). When we mixed the positive and negative control beads, we could distinguish signals from the two different populations and the histograms for the two different populations produced peaks that overlapped only slightly (fig. 4.3 (e)). By contrast, we could not see signal from the T-bet + beads that were washed in PBS prior to antibody addition (fig. 4.3 (f)), suggesting that this washed off the transcription factor bound to the beads. This was most likely because the PBS wash destabilised the interaction between the transcription factor and the oligonucleotide, whereas the oligo buffer was designed to stabilise this interaction. The dissociation constant between a transcription factor and its DNA binding site is generally in the μM range. This is compared to a dissociation constant on the order of $\times 10^{-14}$ for the interaction between biotin and streptavidin. Since, we were trying to avoid having to wash the beads in the first place, this was not an issue.

In flow cytometry, if we do not wash the beads after incubation, this should not overly affect the results as the antibody size is below the limit of detection and will not register as an event. However, free antibody in the buffer will

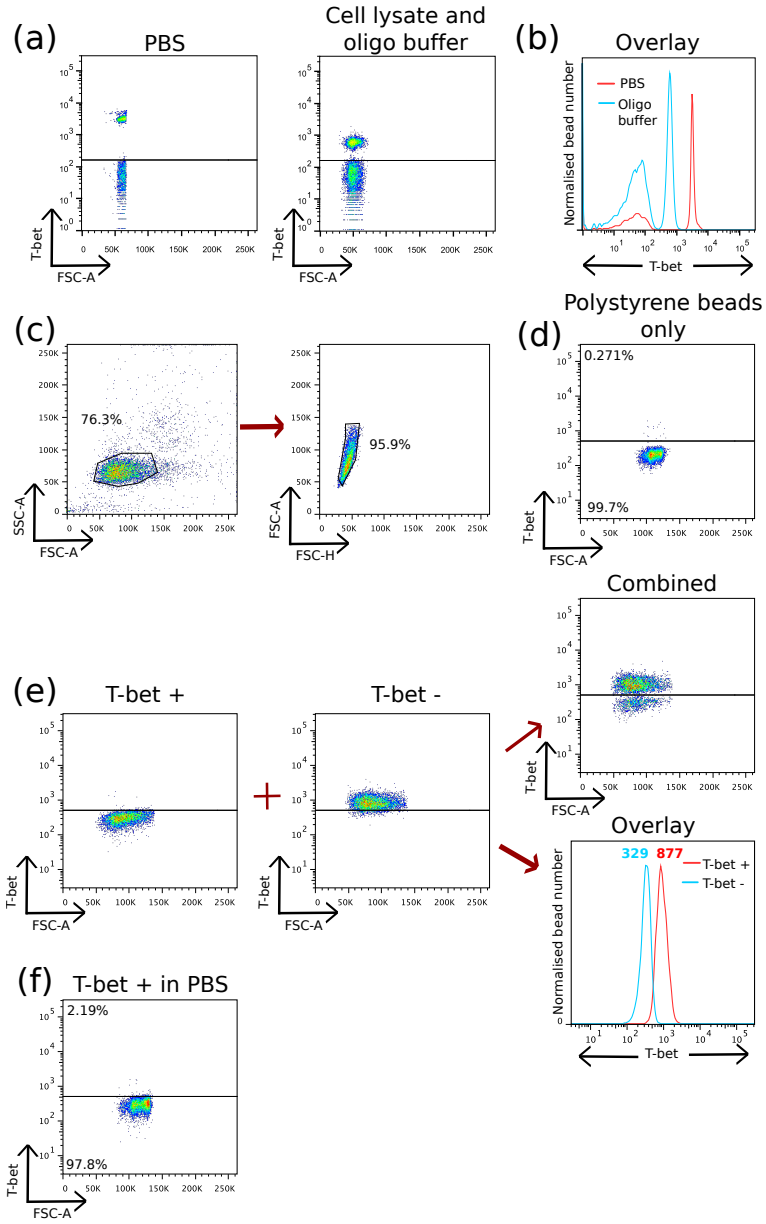

Figure 4.3: Establishing antibody stability in OligoFlow - Positive and negative control T-bet oligonucleotides were used to test antibody performance in buffer used for oligonucleotide pulldown protocol. (a) Plots showing brightness of αT-bet Alexa647 antibody on FacsComp beads in PBS or in cell lysate and buffer used for oligonucleotide pulldown. (b) Overlay of histograms for data shown in (a). (c) Gating strategy for polystyrene beads in OligoFlow experiments. (d) Plot showing fluorescence level of polystyrene beads without addition of antibody. (e) Plots for T-bet on beads made with positive control (T-bet +), negative control (T-bet -), *in vitro* combination of beads for positive and negative control (Combined) and *in silico* overlay of histograms for positive and negative controls (Overlay). Numbers on histogram denote MFI for the correspondingly coloured peak. Beads gated as in (c). Histograms normalised to total events for FSC-A, FSC-H gate. (f) T-bet on beads made with positive control oligonucleotide but washed and resuspended in PBS before incubation with αT-bet antibody.

give a certain level of background fluorescence to all samples. This background should vary less between experimental samples within an experiment than the variation introduced by washing beads in a standard oligonucleotide pulldown assay. However, if the background is too high, it will decrease sensitivity of the experiment. Given that the variations in binding strength caused by a SNP might be quite small we wanted the lowest background (and so most sensitivity) possible while still achieving high signal in the positive control. We were able to achieve this by titrating the antibody dose, which decreased the MFI of the negative control (fig. 4.4 (a)) without affecting the MFI of the positive control (fig. 4.4 (b)).

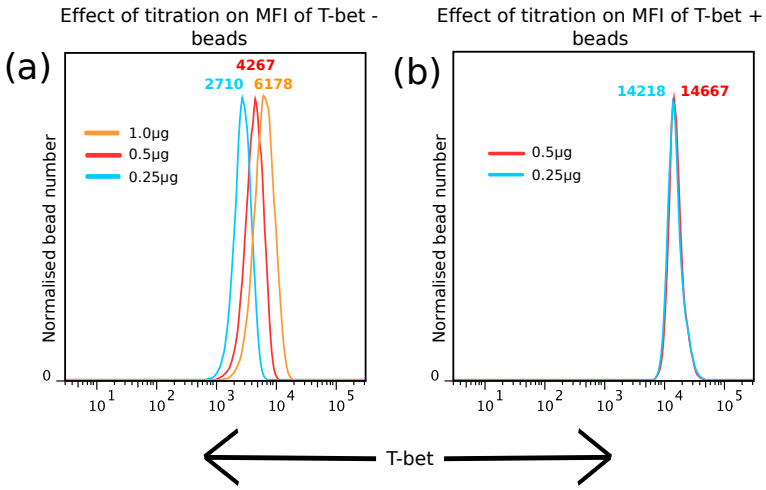

Figure 4.4: Increase in OligoFlow sensitivity by decreasing antibody concentration. - OligoFlow was performed with T-bet + beads and T-bet - beads and decreasing volumes of antibody. (a) Histograms showing MFI of OligoFlow samples using T-bet - beads with 1 μg per test (orange line), 0.5 μg per test (blue line) or 0.25 μg per test (red line) of αTbet antibody. (b) Histograms showing MFI of OligoFlow samples using T-bet + beads with 0.5 μg per test (blue line) or 0.25 μg per test (red line) of αT-bet antibody. Numbers on histogram denote MFI for the correspondingly coloured peak.

We also tested another fluorochrome, eFluro660 which is brighter than Alexa647 but, like Alexa647, is read on the APC channel. We found that this antibody gave decreased sensitivity in this assay because the unbound antibody from the unwashed beads was too bright for the sensitivity required (fig. 4.5).

To get good discrimination between the positive and negative control and thus good sensitivity, we also found that we needed lysate from a high number of cells. Using a high number of cells to prepare the lysate (typically 30×10^6 cells per sample) gave both a lower MFI for the negative control and a higher MFI for

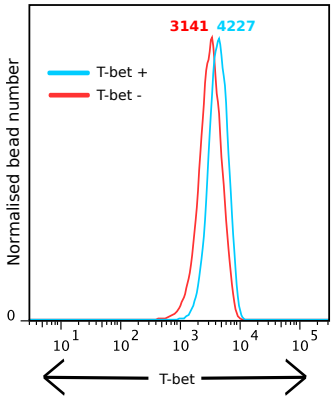

Figure 4.5: Very bright fluorochrome caused high background and low sensitivity in OligoFlow. - We tested T-bet + beads and T-bet - beads in OligoFlow with eFluro660 labelled αT-bet. Histograms for MFI of T-bet+ beads (blue) and T-bet - beads (red) are shown.

the positive control than when the number of cells was lower, thus increasing our dynamic range (fig. 4.6).

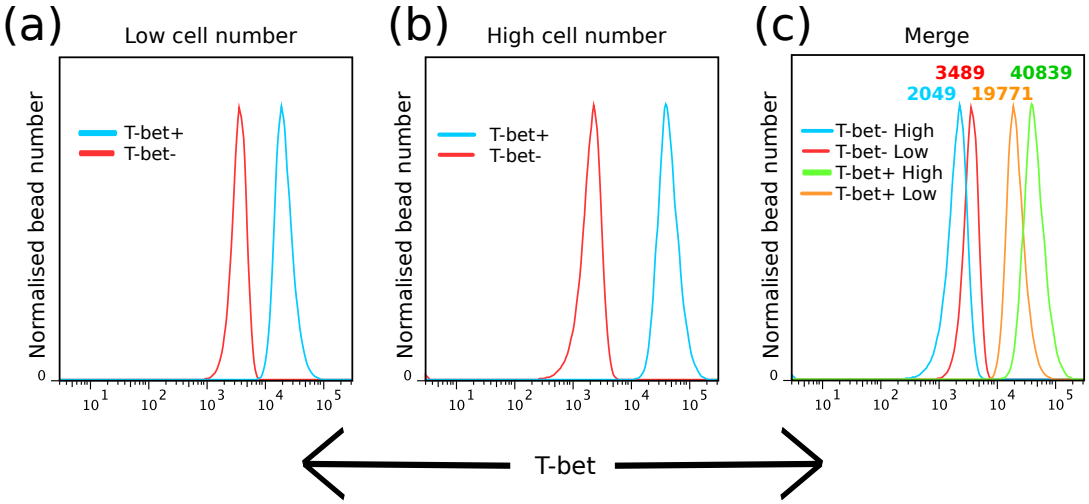

Figure 4.6: High OligoFlow sensitivity with lysate prepared from high cell number. - We prepared lysate from 15×10^6 cells per sample (a) or 30×10^6 cells per sample (b) and used these lysates for OligoFlow with T-bet+ and T-bet- beads. (c) Overlay of MFI of each sample tested. Numbers on histogram denote MFI for the correspondingly coloured peak.

Having optimised the assay on the basis of the positive and negative control oligonucleotides, we were able to run the assay with the rs1465321 oligonucleotides that had shown differential binding in the original pulldown and Western blotting. As before, we found that the SNP rs1465321 showed differential binding between the A and G alleles and this was consistent across multiple

experiments with lysate from both YT cells and $CD4^+$ cells cultured in Th1 conditions. The experiment was repeated with a new lot of oligonucleotides to validate the method and check that the differential binding did not result from any possible issues during the preparation of the original oligonucleotides (fig. 4.7).

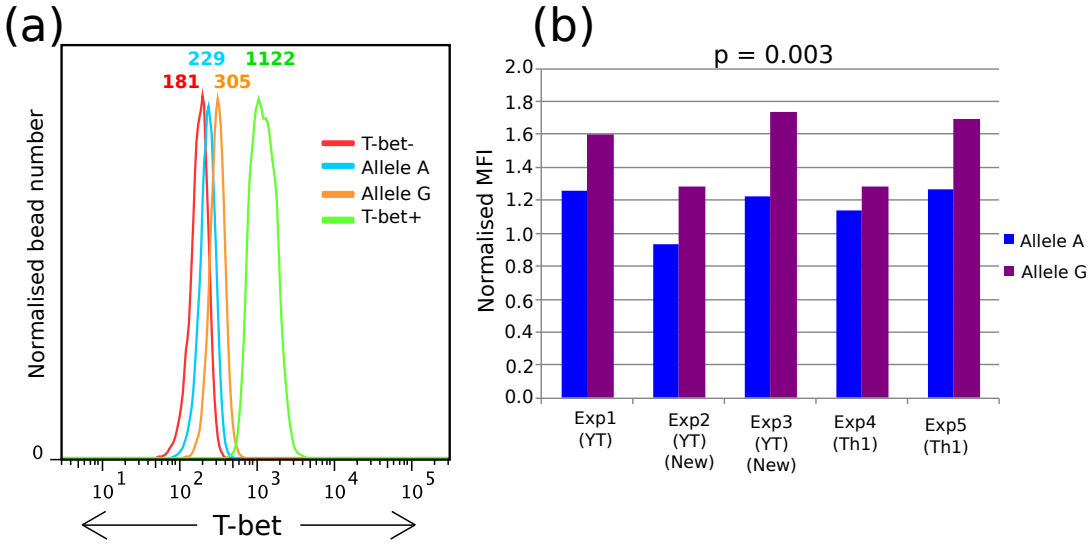

Figure 4.7: rs1465321 alters T-bet binding in OligoFlow assay. - Summary of OligoFlow results for rs1465321. (a) Example plot showing MFI values for T-bet on positive and negative controls and A and G alleles of rs1465321. (b) MFI values for A and G alleles over several experiments (Exp) using YT and Th1 cells and with new batch of rs1465321 oligonucleotide (denoted 'new' under appropriate columns on graph). Values normalised to negative control = 1. p values are for comparison of the two MFI distributions obtained for each allele within each replicate experiment.

4.2.3 Testing other Hit-SNPs

We used the OligoFlow technique to analyse some of the other SNPs that we had found in a T-bet binding site. Since we could not test all the binding site hits from our computational analysis, we chose SNPs based on other data from our *in silico* work. For example, we chose some SNPs because they disrupted the T-bet motif and we might expect that this would impact on binding. However, rs1465321 does not disrupt a T-bet motif - rs1465321 is three base pairs away from a motif, suggesting a slightly more complex situation and implying that we could miss hits if only restricting ourselves to SNPs in a motif. Therefore, we also chose some SNPs that were within 20bp of, but not directly in, a binding

site motif. We also chose SNPs based on the presence of histone modifications or DNase hypersensitivity that might suggest the site was functionally relevant. The SNPs chosen and the reasons for choosing them are outlined in table 4.1. We tested all of these SNPs at least once using lysate from YT cells and/or Th1 cells.

SNPs that did not appear to show differential binding were not tested any further due to time constraints (fig. 4.8). In cases where the first test suggested that there was or might be differential binding, the experiments were repeated with lysate from YT and/or Th1 cells to examine whether a consistent difference could be seen across experiments (fig. 4.9).

Table 4.1: Summary of SNPs chosen for OligoFlow. Hit-SNPs in T-bet binding sites chosen for OligoFlow assay and reasons for choosing them.

Hit-SNP	Associated-Trait	Reasons for Choosing
rs1006353	Body Mass Index	Consensus motif altered
rs10152590	Height	Consensus motif altered. H3K4me3 mark present. DNase sensitive region
rs11135484	Crohn's Disease	Near consensus motif. In intron for *ERAP2* which has a role in MHC peptide loading. H3K4me1 and H3K4me3 marks present. DNase sensitive region.
rs13333528	Colorectal Cancer	Near consensus motif. In intron for *CDH1*. Alters RUNX motif. DNase sensitive region.
rs1420106	Coeliac Disease, Crohn's Disease	Near 5′ end of *IL18RAP* which is involved in signalling in immunity. H3K4me3 mark present. DNase sensitive region.
rs2106346	Psoriasis	H3K4me1 mark present. In intron for *TSC1*
rs2387397	Rheumatoid Arthritis	Consensus motif altered. H3K4me3 mark present. DNase sensitive region.
rs2703078	Coeliac Disease	Consensus motif altered. H3K4me1 mark present. DNase sensitive region.
rs2984920	Coeliac Disease	Near 5′ end of *RGS1*. H3K4me3 mark present. Already tested in pulldown assay and Western blot. Near T-bet consensus sequence.
rs3091310	Coeliac Disease	Within locus for *CCR3*. H3K4me3 mark present. DNase sensitive region.
rs5778	Body Mass	Consensus motif altered. Within 2kbp of *TRHR*.

Hit-SNP	Associated-Trait	Reasons for Choosing
rs6784841	Coeliac Disease	Near consensus motif. In intron for *FRMD4B*. DNase sensitive region.
rs743776	Rheumatoid Arthritis, T1D	Consensus motif altered. Within 50kbp of *IL2RB* which is involved in signalling in the immune system.
rs7441808	Rheumatoid Arthritis, T1D	Consensus motif altered. DNase sensitive region.
rs8008961	Primary Biliary Cirrhosis	H3K4me3 mark present. In intron for *RAD51B*.
rs8062727	Leprosy	Consensus motif altered. H3K4me1 mark present. DNase sensitive region.

Due to the novelty of the assay, there was no pre-defined cut-off point for differential binding. In cases of no differential binding, we would expect the MFI values for the two alleles to be equal and, thus the ratio of the two alleles to be one. Therefore, to test for significance in the differential binding between the A and G alleles of rs1465321, we took the natural logarithm of each MFI value and analysed by paired t-test. This showed the differential binding to be significant ($p = 0.003$). We performed the same analysis on all SNPs tested (table 4.2).

Table 4.2: Significance testing for differential binding in OligoFlow assay. Difference between ln(MFI) values of the two alleles for each SNP were tested and ranked in order of significance.

Hit-SNP	Number of replicates performed	Mean of MFI values	p value
rs1465321	5	0.263	0.003
rs11135484	3	0.172	0.080
rs1006353	3	0.091	0.129
rs2703078	2	0.118	0.172
rs1015290	2	0.055	0.276
rs2984920	2	0.130	0.431
rs13333528	2	0.037	0.450
rs8008961	2	0.023	0.742
rs142016	2	-0.012	0.753
rs2106346	1	0.020	n/a

Hit-SNP	Number of replicates performed	Mean of MFI values	p value
rs2387397	1	0.008	n/a
rs3091310	1	0.033	n/a
rs5778	1	0.013	n/a
rs6784841	1	0.062	n/a
rs743776	1	0.020	n/a
rs7441808	1	0.063	n/a
rs8062727	1	0.039	n/a

By this analysis, only rs1465321 reached significance. This was possibly because we were examining very small differences in binding and three replicates is not enough to find a significant difference. Due to time constraints, we were not able to repeat the assays further. Furthermore, throughout data acquisition, the protocol was optimised to increase the difference between the MFI of the negative and positive controls to try and detect small differences in binding. There was large variability in the MFI of our positive control (T-bet +) between experiments due to variations in number of cells used and efficiency of cell lysis in each experiment. Although we examined the effect of cell number on the experiment, cell number was sometimes suboptimal due to number of cells available. In those experiments where cell number was suboptimal then our experimental range would have been decreased and we would expect to see smaller differences between two alleles of any 'hit' SNP for technical reasons. Likewise, in conditions of ample lysate, we might see small experimental variations amplified enough to give a false positive. In addition to increasing cell number and optimising antibody concentration, as already discussed, we also varied the voltage for the fluorescence channel. As such, there is wide inter-assay variability in the experiments which is unideal for comparing MFI values across experiments. Now that the protocol is established, moving it towards a robust medium to high throughput assay might include the standardisation of settings across all experiments using Cytometer Setup and Tracking beads to set a standard baseline on the cytometer, combined with a fixed voltage in the fluorescence channel. This would allow cleaner comparison across replicate experiments for the same SNP.

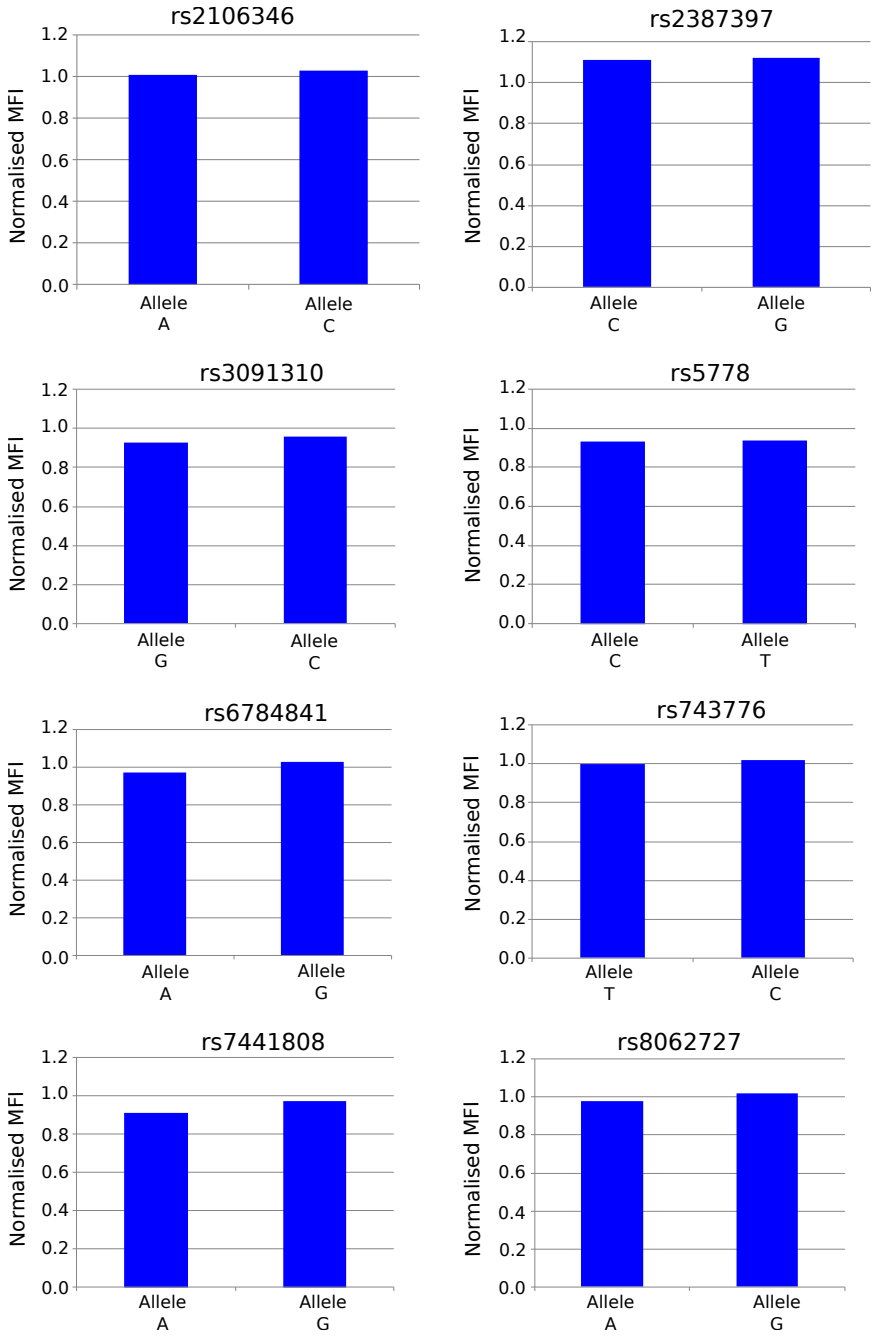

Figure 4.8: Some hit-SNPs did not show differential binding in OligoFlow. - MFI values normalised to the negative control (T-bet -) = 1 are shown for those hit-SNPs that did not show differential binding in OligoFlow.

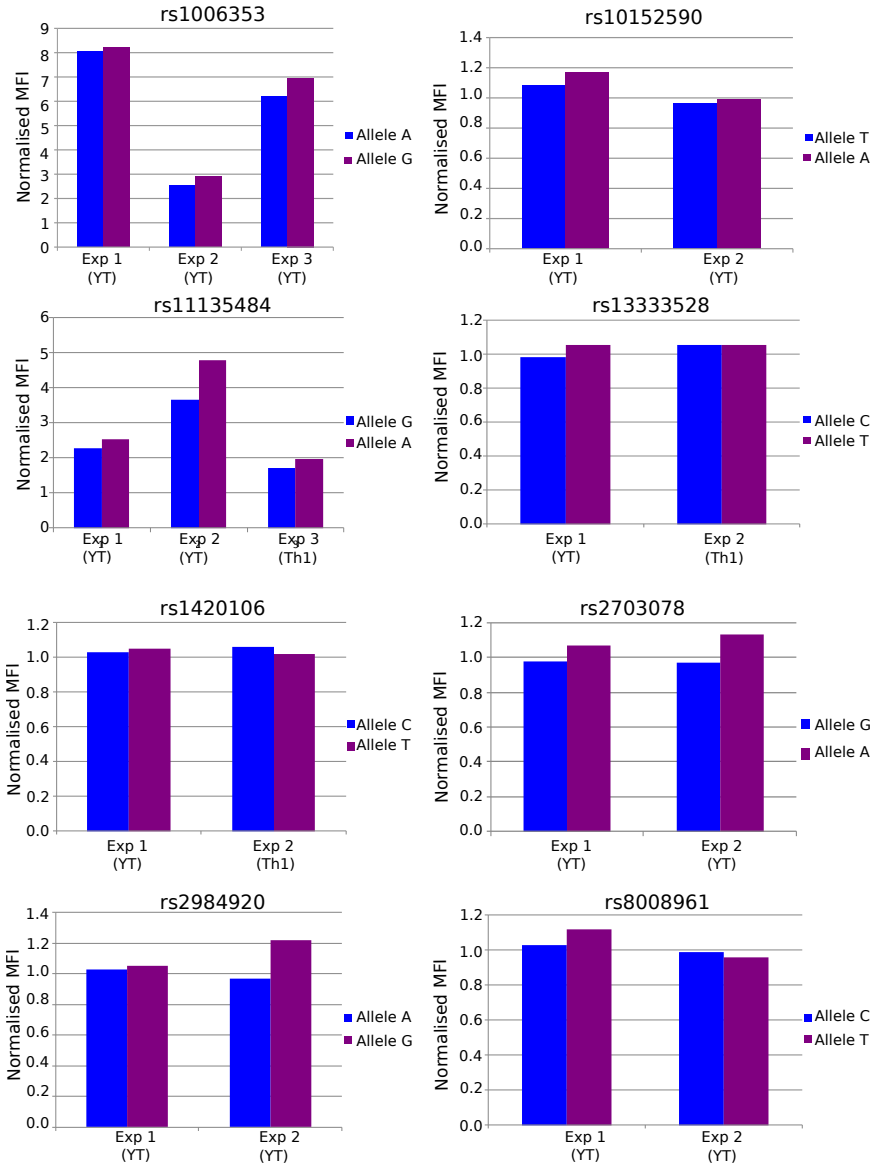

Figure 4.9: Some hit-SNPs were tested more than once to search for consistent differential binding. - MFI values normalised to the negative control (T-bet -) = 1 for hits which showed possibility of differential binding in first OligoFlow assay and were repeated. Each pair of bars represents one experiment (Exp). YT denotes experiment performed with lysate from YT cells. Th1 denotes experiment performed with lysate from *in vitro* cultured human Th1 cells.

Moreover, this assay was a screening step and it would be more biologically relevant to test for altered binding of these hit-SNPs from the OligoFlow assay *in vivo*. To gain some idea of where a cut-off for differential binding might lie we plotted the difference in ln(MFI) values (fig. 4.10). This suggests that a cut-off of approximately 0.1 for ln(MFI of allele1) - ln(MFI allele2) would be appropriate. However, more work is needed to validate this, including repetition of rs1006353, rs11135484 and rs2703078 to check whether the differences for these are significant if we perform more replicates. Moreover, the results of the assay need validation by a further method such as ChIP followed by allele-specific qPCR. Based on the results from the OligoFlow, we took our top three hits rs1465321, rs111354854 and rs1006353 to be of interest in further analysis and discussion but with the above caveats.

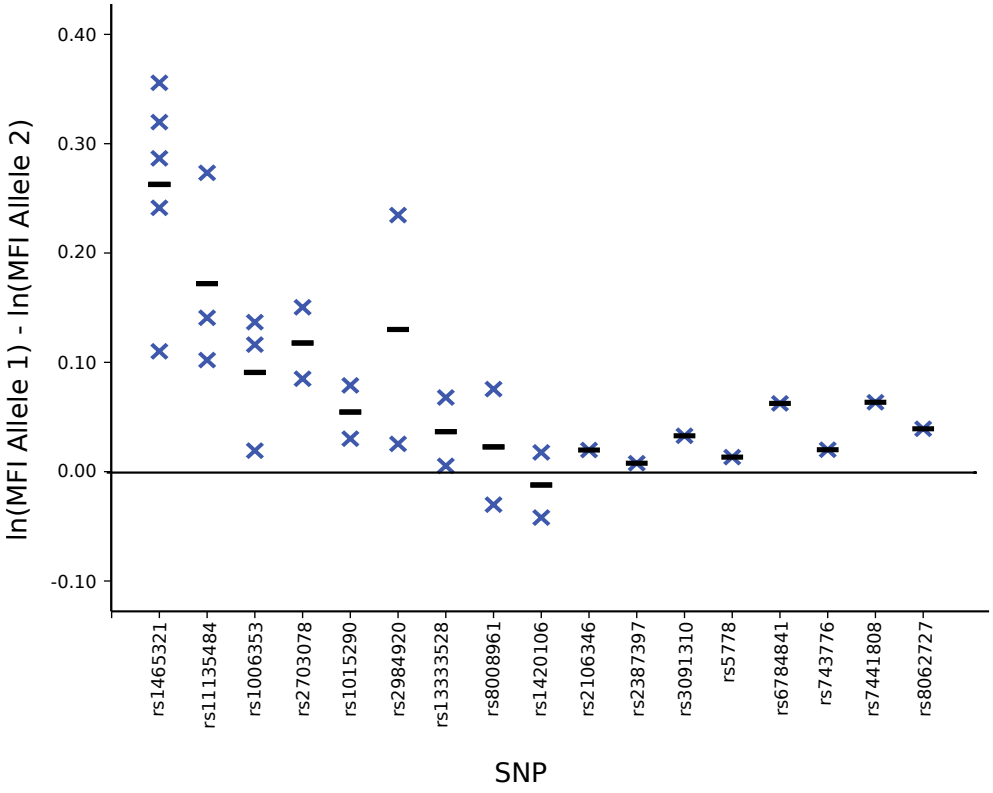

Figure 4.10: Difference in ln(MFI) values for all replicates of all SNPs tested. - Natural logarithm of each MFI value was taken and difference between the two alleles for each experiment is shown. Data is as used in table 4.2. Difference of 0 (MFI values the same) is shown as solid line across graph.

We tried as well to test some of the binding site hits that we had found for GATA3 by OligoFlow. We made positive and negative control oligonucleotides as we had for T-bet and incubated these with lysate from the Jurkat cell line, which expresses GATA3. However, we found that we could not get the same level of separation between positive and negative controls as we could for T-bet, just by changing two residues. We tried two separate positive and negative control pairs (V1 and V2, fig. 4.11) but found that, although we could obtain a range of MFI values across the two pairs, the MFI values of each pair depended heavily on the surrounding sequence (which differed between V1 and V2 and is given in materials and methods table 2.3). The MFI values between the positive and negative control of each pair showed a small separation in MFI (fig. 4.11 (c)). This may demonstrate the relative promiscuity of GATA3 binding relative to DNA sequence versus T-bet and raises questions over whether the assay could work to assay single base pair changes. However, work is ongoing.

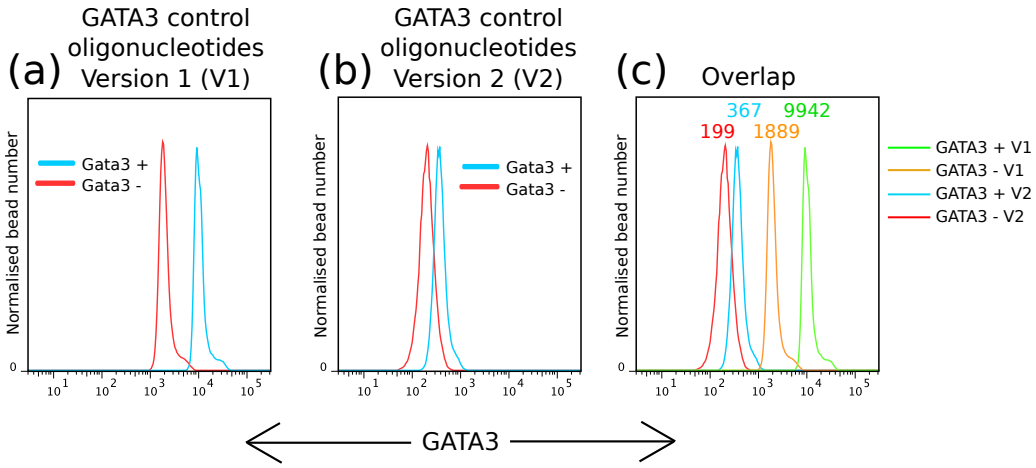

Figure 4.11: Separation for GATA3 positive and negative controls. - Oligonucleotides containing either the complete GATA3 consensus motif (Gata3+) or the GATA3 consensus motif with two residues mutated (Gata3-) were designed. (a) and (b) each show a pair of control oligonucleotides where Gata3+ and Gata3 - differ by only two base pair changes in the consensus motif. (a) shows result for version 1 of oligonucleotides (as labelled in materials and methods). (b) shows results for version 2 oligonucleotides. Surrounding sequence for the consensus motif differs between (a) and (b) and shifts the MFI of the pair as can be seen by overlapping results for the two pairs (c).

4.3 Further *In Silico* Analysis of *In Vitro* Hits

4.3.1 rs1465321

One of our biggest 'hits' from the binding studies was a SNP in the second intron of the *IL18R1* gene. This SNP, rs1465321, is in an LD block associated with coeliac disease[214], Crohn's disease[319] and, to a lesser extent, ulcerative colitis.[279] T-bet binds across this locus both at *IL18R1* and *IL18RAP* and, indeed, our *in silico* data also showed another potential binding site SNP, rs1420106 just upstream of *IL18RAP*. GATA3 also binds across this locus. The SNP rs1465321 is also in a binding site for GATA3 in Th1 cells and rs1420106 is in a binding site for GATA3 in both Th1 and Th2 cells. Furthermore, two SNPs in between the *IL18R1* and *IL18RAP* genes are also in a GATA3 binding site in both Th1 and Th2 cells (though they are not in a T-bet binding site.) These SNPs are rs12991737 which is in LD with rs3771180, a SNP associated with Asthma[320] and rs3732123 which is in LD with rs17027258 a SNP associated with white blood cell subtypes[321]. Another SNP rs2058622, also found in a GATA3 binding site in both Th1 and Th2 cells, is found downstream of rs1465321, in the same intron of *IL18R1*. Although not itself in a called T-bet binding site, rs2058622 sits between two called T-bet peaks in a region of high T-bet binding. The rs2058622 SNP is in the same LD block as rs1465321 and is also, therefore, associated with coeliac and Crohn's disease and ulcerative colitis. Alignments of T-bet and GATA3 binding in addition to histone marks and regions of DNase hypersensitivity across the locus are shown in figure 4.12. T-bet also binds across the region in mouse both in Th1 cells and Treg cells. Our lab also has data on the binding of another master regulator, FoxP3, in the Treg cell lineage in mouse cells and this also shows binding across the locus (figure 4.13.)

To investigate whether anything was already known about correlation of our SNP with gene expression we checked the original papers for the disease-associated SNP as lodged in the NHGRI GWAS catalogue and also searched for our SNPs in the Gene Expression Variation (Genevar) tool from the Wellcome Trust Sanger Institute. The SNP rs1465321 is in LD with three published GWAS hits rs13015714, rs2058660 and rs917997. In the original papers, the haplotype at both rs13015714 and rs917997 were shown to be associated with expression of IL18RAP but not IL18R1 in whole blood in coeliac cases.[214] Genevar can be used to analyse 'in-house' data but also provides published information

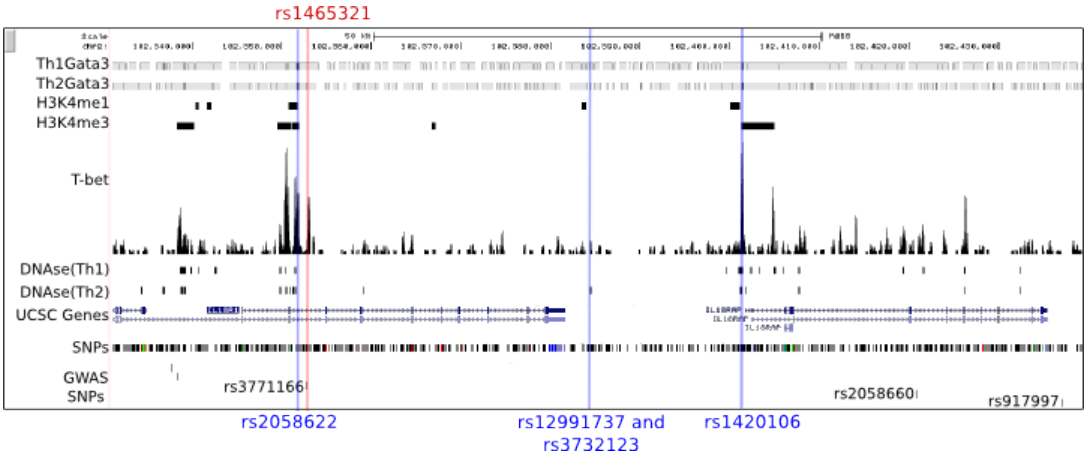

Figure 4.12: Genomic annotation of region surrounding rs1465321. - UCSC genome browser plot showing T-bet binding at the *IL18R1/IL18RAP* locus. GATA3 binding in Th1 cells and Th2 cells and called peaks for histone marks and DNase hypersensitivity as used in results chapter one are also shown. Bottom track shows location of other SNPs in the region that are trait-associated SNPs in the NHGRI GWAS catalogue. The first and second SNPs (from left to right) in the this track are not labelled to maintain clarity of figure but are rs9807989 and rs13015714 respectively. The location of rs1465321 is shown as a red line across the figure. The locations of other binding site SNPs are shown as blue lines across the figure. Because rs12991737 and rs3732123 cannot be resolved at the scale shown, they are indicated by one line.

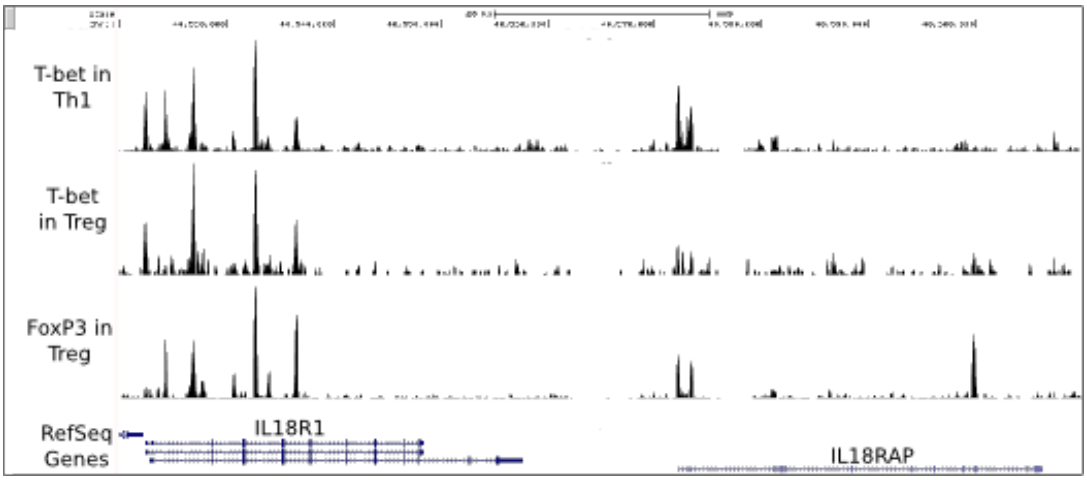

Figure 4.13: T-bet and FoxP3 binding at *Il18r1/Il18rap* in mice - UCSC genome browser plot showing T-bet binding at the *Il18r1/Il18rap* locus in Th1 and Treg cells in mouse and FoxP3 binding in Treg cells across the locus in mouse.

on eQTLs from cell lines from the HapMap3 project,[322] the Gencord collection[46] and the MutHER project.[323] These projects contain a variety of cells lines immortalised from various cell types (including T cells, B cells and non-immune cells) originally donated from a variety of people across a variety of ethnicities. However, neither rs1465321 nor rs142016 were present in the database for this resource.

When we looked back at the TRAP data for the sequences around each of our binding site SNPs, we found that the sequence around rs1465321 was suggested to have potential binding sites for AP1, FEV, ELV5, NFE2L2 and RORA_2. TRAP also has a tool for predicting whether a SNP alters the binding of a transcription factor in the JASPAR database. This tool did not return any significant hits but it is important to remember that the JASPAR database is not exhaustive and does not, for example, include a PWM for T-bet.

4.3.2 rs11135484

The SNP rs11135484 is in an intron of the *ERAP2* gene and is in LD with rs2549794, a SNP associated with Crohn's disease.[319] It is within a binding site for T-bet but also GATA3 in both Th1 and Th2 cells (fig. 4.14) and is in LD with rs2549794, a SNP associated with Crohn's disease[319]. As its full name, Endoplasmic Reticulum Aminopeptidase 2 suggests, ERAP2 is involved in peptide trimming: it is required for the generation of most MHC class I peptides in human but is not found in mouse.

When we searched for eQTL data using Genevar, we found that rs11135484 was associated with altered ERAP2 expression in the HapMap collections (figure 4.15). The TRAP data suggested that the sequence around rs11135484 could have potential binding sites for Runx1, SpI1, IRF1, FEV, ELV5 and IRF2 and that the SNP could interfere with an Egr1 binding site.

4.3.3 rs1006353

As seen in results chapter one, rs1006353 is not within 2kbp of a gene. It is, however, within 50kbp of Mitochondrial Translational Initiation Factor 3 (MTIF3) and General Transcription Factor 3A (GTF3A). As can be seen from figure 4.16,

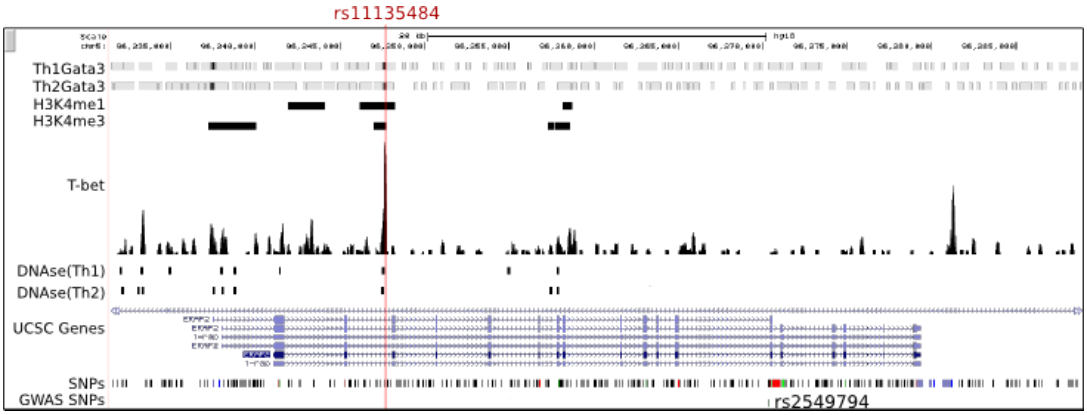

Figure 4.14: Genomic annotation of region surrounding rs11135484. - UCSC genome browser plot showing T-bet binding at the ERAP2 locus. GATA3 binding, methylation marks and DNase hypersensitivity as used in analysis in results chapter one is also shown. The location of rs11135484 is shown as a red line across the figure.

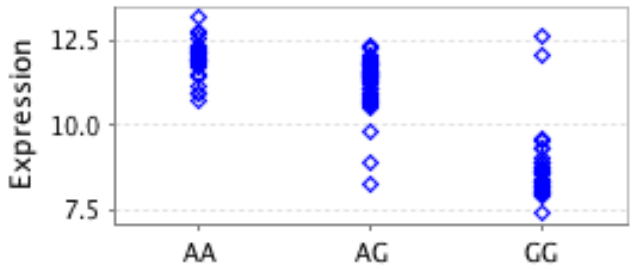

Figure 4.15: rs11135484 is an eQTL. - Expression of ERAP2 by genotype at rs11135484 in lymphoblastoid cell lines from HapMap 3 project as analysed by Genevar.

it is in a binding site for T-bet but also for GATA3 in both Th1 and Th2 cells. Interestingly, it is also approximately 70kbp from *LNX2* an E3 ligase which regulates the cellular location of CD8α.[324] Another SNP, rs17753121, which is also found between the genes for MTIF3 and LNX2, is in a GATA3 binding site in Th2 cells. The SNP rs17753121 is associated with obesity while rs1006353 is in LD with rs4771122, a SNP associated with Body Mass Index in a separate study.[325]

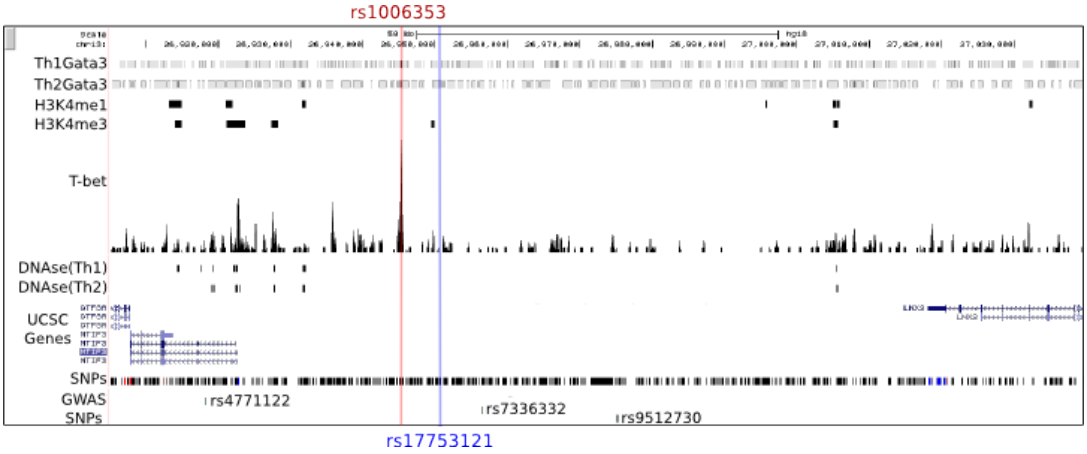

Figure 4.16: Genomic annotation of the region surrounding rs1006353 - UCSC genome browser plot showing T-bet binding around rs1006353. GATA3 binding, methylation marks and DNase hypersensitivity as used in analysis in results chapter 1 is also shown. The location of rs1006353 is shown as a red line across the figure. The location of rs17753121, a SNP in a GATA3 binding site in Th2 cells is shown as a blue line across the figure.

Although found in the Genevar databases, rs10006353 was not significantly associated with gene expression as an eQTL. In the original study, some association was found between rs4771122 and expression of GTF3A in lymphocytes and blood but this was best explained by another SNP, rs7988412, which is also in LD with rs4771122. The SNP rs7988412 is located on the other side of the trait-associtaed SNPs rs4771122 to our hit-SNP, rs1006353, on the chromosome. Our TRAP analysis suggested potential binding sites for Runx1 and SpI1 in the 200bp region around rs1006353. Reflecting the gained GATA motif, as shown in results chapter one, our TRAP analysis suggested that the SNP may alter GATA1 binding. The analysis also suggested the SNP may interfere with an Evi1 motif.

4.4 Functional Testing of OligoFlow hits

We wanted to test whether hit-SNPs that showed differential binding in our OligoFlow assays would show differential gene activation in a luciferase assay. Many of our hit-SNPs were outside of promoter regions. We reasoned that the regions containing these SNPs must exert their effect by acting on a promoter region and that we would need to include a promoter region in our luciferase constructs in order to see any effect from differential binding at the SNP under investigation. Since the best know promoter target of T-bet is the *IFNG* promoter, we decided to make a general construct with this promoter upstream of the firefly luciferase gene into which we could then insert a region containing a SNP of interest from the OligoFlow assay. Therefore, we cloned the promoter region from -456 to +86 relative to the *IFNG* TSS into the PGL4.13 firefly luciferase construct. (The PGL4.13 vector contains an SV40 promoter but the *IFNG* promoter was inserted in such a way as to disrupt this promoter so that only the *IFNG* promoter was controlling luciferase expression.) The region from -456 to +86 was chosen based on the -445 to +64 construct used by Soutto *et al* and includes the -445 to -410 region shown to be critical for T-bet responsive promoter activation by that group.[180] We cotransfected this construct into EL-4 cells with a renilla luciferase transfection control and tested for firefly luciferase expression that was increased when a T-bet expressing plasmid was also cotransfected (fig. 4.17).

Of all the SNPs tested by OligoFlow, rs1456321 showed one of the most consistent and biggest differences in T-bet binding. It is also not in a promoter region. Therefore, we decided to test this SNP in our luciferase construct. We cloned a region of 447 bp around rs1465321. The size of this region was designed to be large enough to try and recapitulate some of the genomic environment around rs1465321 but small enough not to include any other common SNPs as determined from HapMap and from pilot phase 1000 genome data. We made two separate constructs, one with the A allele at rs1465321 and one with the G allele. We cotransfected this construct into EL-4 cells with a plasmid that expressed T-bet or an empty expression plasmid. We also transfected a renilla luciferase expressing transfection control into all samples. After 24 hours, we stimulated the cells and assayed for luciferase expression. To control for transfection efficiency we divided firefly luciferase expression by renilla luciferase expression. To standardise across experiments, we normalised each experiment

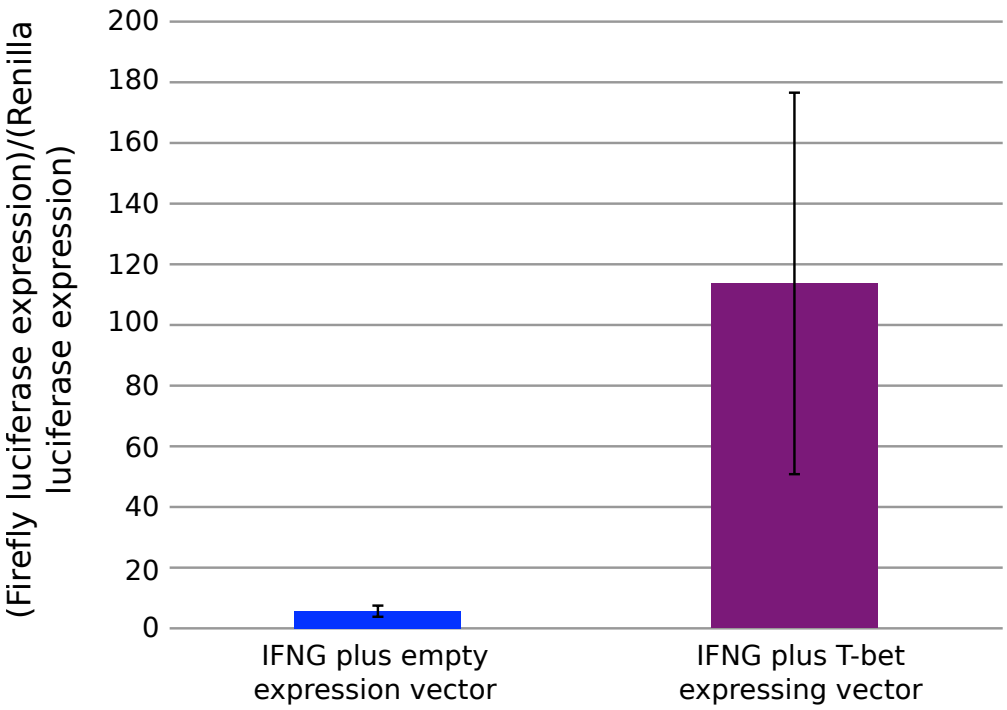

Figure 4.17: T-bet responsive luciferase construct. - Firefly luciferase expression normalised to renilla transfection control from *IFNG* promoter construct when co-transfected with T-bet expressing plasmid or empty expression vector control. Results are averaged over three separate experiments. Error bars show standard deviation across three different experiments.

to (firefly/renilla) expression for IFNG promoter alone =1. We found a slight increase in firefly luciferase signal from the construct containing the G allele of rs1465321 versus the construct containing the A allele (fig. 4.18) and this was consistent across three experiments. This was not significant (p = 0.061, 2 tailed paired t-test), most likely because we are looking at very small differences in signal across only three experiments. Because of the small differences, we would need to further repeat the assay before we could draw any firm conclusions and it is possible that the luciferase assay system introduces too much experimental noise to assay the differences in binding properly. Furthermore, while the region around rs1465321 does seem to act to enhance transcription of the luciferase gene slightly, it does not show the strong enhancer activity typically seen in luciferase assays. This may imply that this particular region is not, in fact, functionally relevant. However, the presence of the T-bet binding site, in addition to other genomic marks as seen in our computational analysis, would suggest otherwise. Alternatively, the region may interact with other parts of the *IL18R1* locus to exert its full biological function and this is missed in a luciferase assay that extracts the region from its full genomic context.

4.5 Conclusion

We have found three SNPs that are in a binding site for T-bet, alter binding of T-bet *in vitro* and are in high LD with SNPs associated with coeliac disease, Crohn's disease, ulcerative colitis or Body Mass Index. These SNPs are in regions of genomic and immunological importance. We have developed a luciferase construct to test the functional effect of altered binding at non-promoter hit-SNPs. We found increased binding of T-bet at one of our hit-SNPs, rs1465321. However, we did not find this to be significantly associated with increased transcriptional activity at this time. This may indicate issues with our approach or suggest that rs1465321 has no causal effect. Alternatively, it could, indicate that the luciferase assay is not the most appropriate tool for assaying such small differences in binding. Recent work from the ENCODE project found that luciferase assays are more useful for verifying transcription factor function at regions close to the TSS. Presumably, binding is more dependant on sequence close to the TSS and is less dependant on the other chromatin elements that are important at more distal sites.[27] Our luciferase results may re-

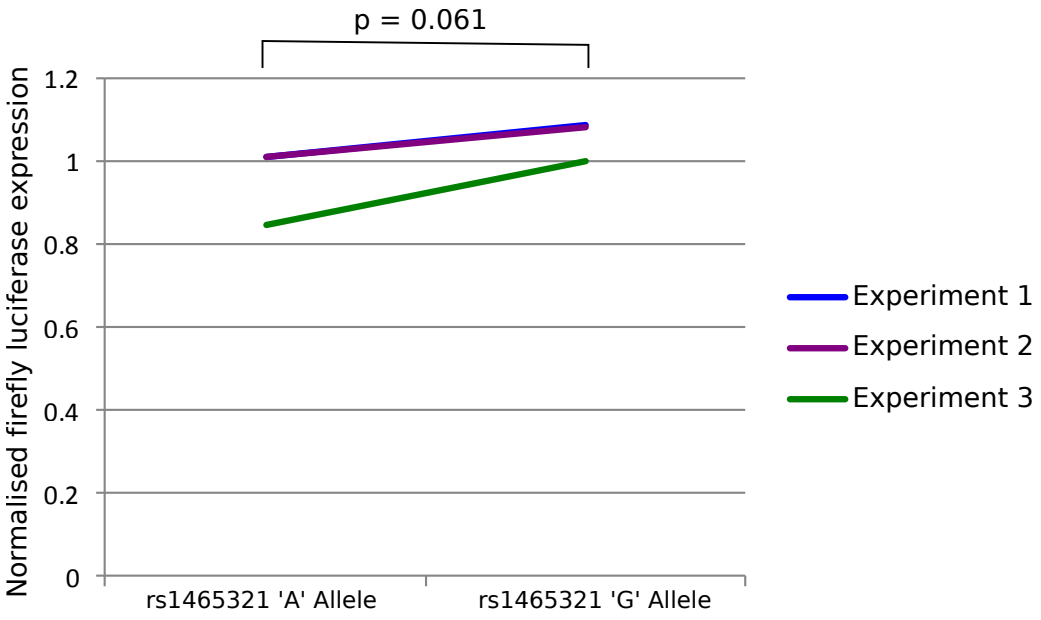

Figure 4.18: Small but non significant increase in luciferase expression with G versus A allele of rs1465321. - Expression of firefly luciferase from constructs containing region around rs1465321 upstream of *IFNG* promoter. Expression is normalised to renilla transfection control and then *IFNG* promoter alone = 1. Results are shown for three separate experiments. Lines connect results from each allele for one experiment. Of note, data for experiments 1 and 2 were very similar and lines overlap on graph.

flect the same issue and further work needs to be done before fully confirming or discounting an effect at rs1465321 or any other of our hit-SNPs.

5

Functional Studies of the IL-18 Receptor

5.1 Introduction

One of the largest and most consistent hits from our *in silico* and *in vitro* work was a SNP, rs1465321, in an intron of *IL18R1*. This SNP is in a binding site both for T-bet and for GATA3 in Th1 cells and is found near a motif for T-bet. FoxP3 also binds at this region in mouse Treg cells: we can hypothesise that the same is true in human Treg. The gene for IL18R1 neighbours the gene for IL18RAP and disease-associated SNPs are found across the region. IL-18 signalling has a key role in Th1 function: IL-18 stimulation can induce production of IFN-γ, the canonical Th1 cytokine and full IL18R1 expression is, at least in part, dependant on T-bet. Despite its importance in immunology, data on the kinetics of IL18R1 and IL18RAP expression is conflicting on some key points. There is disagreement over the level of IL18R1 expression on unstimulated cells and over the extent to which T-bet is required for IL18R1 expression under different circumstances. If we are to functionally connect altered transcription factor binding to downstream mechanistic effects and downstream disease risk then we need a better understanding of the fine details of IL18R1 expression and the role of IL-18 in disease mechanisms. To try and examine some of these fine details, we moved to mouse studies. The *IL18R1/IL18RAP* locus is highly conserved between mouse and human and assays in mice have two important advantages. Knockout models including the T-bet knockout mouse, are available and we

can induce and manipulate disease mechanisms in such studies.

5.2 Kinetics of IL18R1 Expression

5.2.1 IL18R1 Expression on Wild-Type and T-bet$^{-/-}$ Cells

The kinetics of IL18R1 expression have been followed both in C57BL/6 and BALB/c mice by Yu *et al*.[204] This group show that intermediate levels of IL18R1 are expressed on unstimulated CD4^{+} cells and that this expression is then downregulated on activation. IL18R1 is then reexpressed or further downregulated as a cell commits to the Th1 or Th2 lineage respectively. At the mRNA level, Yu *et al* observe co-regulation of IL18R1 and IL18RAP during the process of lineage commitment. However, Yu *et al* use the entire CD4^{+} population, as isolated by magnetic beads, for their starting population. Therefore, they cannot exclude effects of effector and memory cells in the population. As for the role of T-bet, in previous reports the absence of T-bet either diminishes or completely abrogates the upregulation of IL18R1 in the Th1 condition.[170,213]

To determine whether T-bet deficient cells can upregulate IL18R1 at all and to examine the kinetics of IL18R1 expression relative to CD3^{+} stimulation in a pure population of naïve cells, we sorted naïve cells from wild-type and T-bet $^{-/-}$ mice to very high purity. Sorting was based on CD62L^{hi} and CD44^{lo} expression which defines a naïve population.[326] The ligand of CD62L is expressed on the high endothelial venules of peripheral lymph nodes: the expression of CD62L on naïve cells targets them to the peripheral lymph nodes where they may meet their cognate antigen. CD44 is also involved in cell migration and its expression on CD4^{+} cells is a mark of an effector/memory phenotype. We then cultured these cells in Th1 or Th2 conditions and examined the kinetics of IL18R1 expression. We found that, even starting from this pure population, the wild-type Th1 cells upregulated IL18R1 over the length of the timecourse. By contrast the Th2 cells rapidly downregulated IL18R1 such that expression of this molecule was never seen even at early timepoints (fig. 5.1). We found that IL18R1 upregulation was defective in T-bet deficient cells: expression of IL18R1 was noticeably reduced in T-bet$^{-/-}$ Th1 cells versus wild-type Th1 cells at all timepoints (fig. 5.1). However, expression of IL18R1 was higher on T-bet$^{-/-}$ cells cultured in Th1 conditions than either wild-type or T-bet$^{-/-}$ cells in Th2 conditions (figs.

5.1 and 5.2). This result suggested, in contrast to Thieu *et al*,[213] that T-bet $^{-/-}$ cells can express some IL18R1 when cultured under Th1 conditions. Of interest was a small reduction in IL18R1 between 24 and 48 hrs post-activation in the T-bet$^{-/-}$ cells unlike the wild-type cells, where the proportion of IL18R1 positive cells increased with each timepoint. This could illustrate the instability of IL18R1 expression in the absence of T-bet and may suggest why some reports find IL18R1 in T-bet $^{-/-}$ cells and some do not depending on timepoint of assay. Of note, samples from 0 and 12 hours post activation were acquired during the same flow cytometry session. Samples from 24 and 48 hours post activation were also acquired during the same session as each other but during a different session to those samples for 0 and 12 hours post activation. Therefore, care must be taken in directly comparing data from 12 hours and 24 hours. However, gating with respect to the unstained samples in both sessions is similar allowing broad trends to be noted.

In light of our timecourse results we decided to test whether IL-18 could signal in T-bet $^{-/-}$ cells. We cultured T-bet$^{-/-}$ cells in Th1 conditions for six days and then washed the cells and incubated them for 90 minutes in serum free medium. We then added IL-18 to the cultures and measured response to IL-18 by phosphorylation of p38. Phosphorylation of p38 is a downstream effect of IL-18 signalling in CD4^{+} cells.[199] Both wild-type and T-bet$^{-/-}$ cells were able to respond to IL-18 by phosphorylating an already available supply of p38 (fig. 5.3).

5.2.2 IL18R1 Expression on Wild-Type IFN-$\gamma^{-/-}$ Cells

T-bet acts to stabilise the expression of IFN-γ, the production of which is reduced in T-bet deficient cells. We examined whether the reduced capacity for IL18R1 expression in T-bet$^{-/-}$ cells might be mediated through the resultant reduction in IFN-γ expression rather than the loss of T-bet itself. To do this, we assessed IL18R1 expression in IFN-γ deficient mice. We found reduced IL18R1 expression in IFN-$\gamma^{-/-}$ mice compared to a parallel wild-type control but substantial IL18R1 was still seen in these cells (71% positive cells in IFN-γ $^{-/-}$ cells versus 95.8% in wild-type, fig. 5.4). Therefore, we can conclude that IFN-γ is not essential for IL18R1 upregulation. This contrasts with work by Smeltz *et al*[206] which suggests that IFN-γ is required for the induction of high level IL18R1 ex-

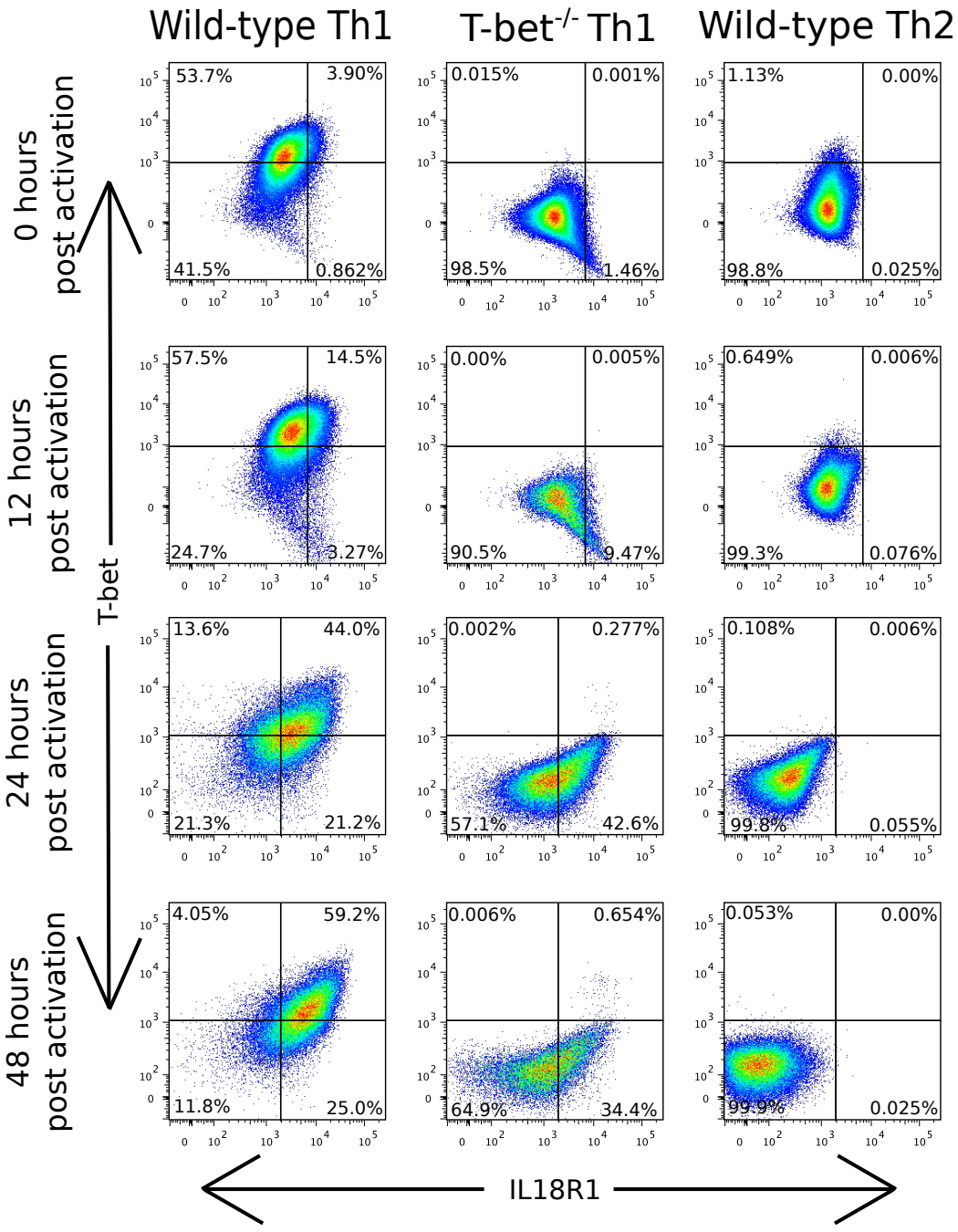

Figure 5.1: T-bet deficient cells are impaired but not unable to express IL18R1 in Th1 conditions. - Naïve $CD4^+$ cells were sorted from wild-type and T-$bet^{-/-}$ mice and activated using αCD3 and αCD28 in Th1 or Th2 conditions. IL18R1 expression was assessed by flow cytometry at varying timepoints after removal from activation.

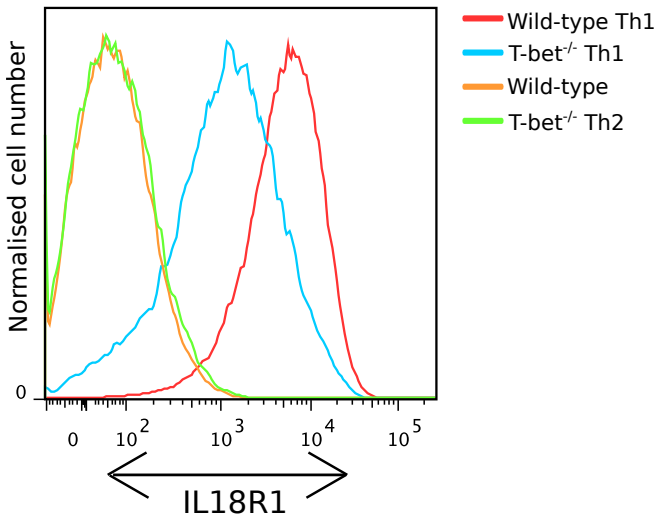

Figure 5.2: T-bet deficient cells have deficient upregulation of IL18R1. - Histograms of IL18R1 expression on wild-type and T-bet$^{-/-}$ cells that have been activated in Th1 or Th2 conditions and cultured for 48 hours after removal from activation.

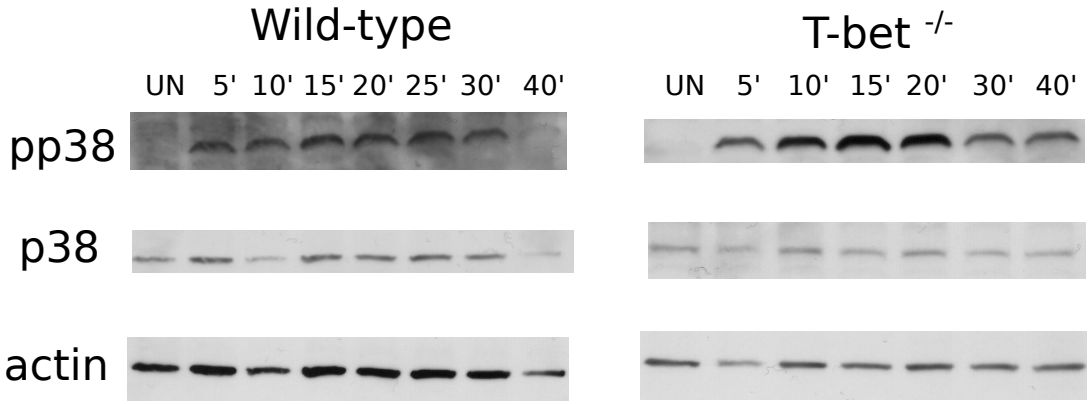

Figure 5.3: T-bet deficient cells can phosphorylate p38 in response to IL-18. - Cells from wild-type and T-bet$^{-/-}$ mice were cultured in Th1 conditions, serum starved and then IL-18 was added for varying lengths of time. Cells were harvested and signalling response was measured by Western blot for pp38. UN denotes unstimulated condition in which cells were serum starved but did not receive IL-18. Numbers denote time in minutes between addition of IL-18 and harvest of cells.

pression by IL-12. However, although Smeltz *et al* do find a substantial reduction in IL-12 driven IL18R1 expression in IFN-$\gamma^{-/-}$ cells they do observe some IL18R1 expression in these conditions. The differences in our data most likely reflect differences between our experiments. Smeltz *et al* used splenic $CD4^+$ cells, activated these for 48 hours and then removed them from activation and cultured for a further 48 hours. We used $CD4^+$ cells that had been sorted naïve and activated these for a longer period of time prior to running our timecourse. Furthermore, Smeltz *et al* do not consider a role for T-bet and do not attempt to separate out the direct effects of IFN-γ loss versus the indirect effects of IFN-γ loss through reduced T-bet expression. Our data supports the idea that IFN-γ is less necessary than T-bet for stable IL18R1 expression given that IFN-$\gamma^{-/-}$ cells reach 71% IL18R1 positive cells versus a maximum of 42.9% in the T-bet $^{-/-}$ cells. However the timecourse experiments are not directly comparable as the IFN$g^{-/-}$ was on a BALB/c background. From Yu *et al*,[204] we would expect BALB/c and C57BL/6 strains to have similar kinetics of IL18R1 expression and so this should not affect the results substantially.

To better address the direct contribution of T-bet and IFN-γ to IL18R1 expression, we used data generated by Jenner *et al*[158] on mRNA levels in T-bet$^{-/-}$, IFN-$\gamma^{-/-}$ double deficient cells that were then transduced with constitutive T-bet before assaying mRNA levels by array (stored at ArrayExpress, Accession E-TABM-759). Transduction of T-bet into these cells was able to upregulate IL18R1 and IL18RAP expression even though IFN-γ was still absent, demonstrating the independence from IFN-γ of T-bet in IL18R1 and IL18RAP expression (fig. 5.5).

5.2.3 IL18R1 and IL18RAP Expression Across $CD4^+$ Subsets and Mouse Strains

IL-18 signals through a heterodimer consisting of both IL18R1 and IL18RAP. There is conflicting data as to whether and under what conditions these two proteins are co-regulated. To assess relative levels of IL18R1 and IL18RAP in different $CD4^+$ cell subsets, we used data generated in house for mRNA levels across Th1, Th2, Th17 and Th0/Thp subsets in BALB/c and C57BL/6 mice. Cells had been sorted as naïve, polarised for seven days and then reactivated with αCD3 and αCD28 for four hours. It should be noted that Balb/c Th0 cells

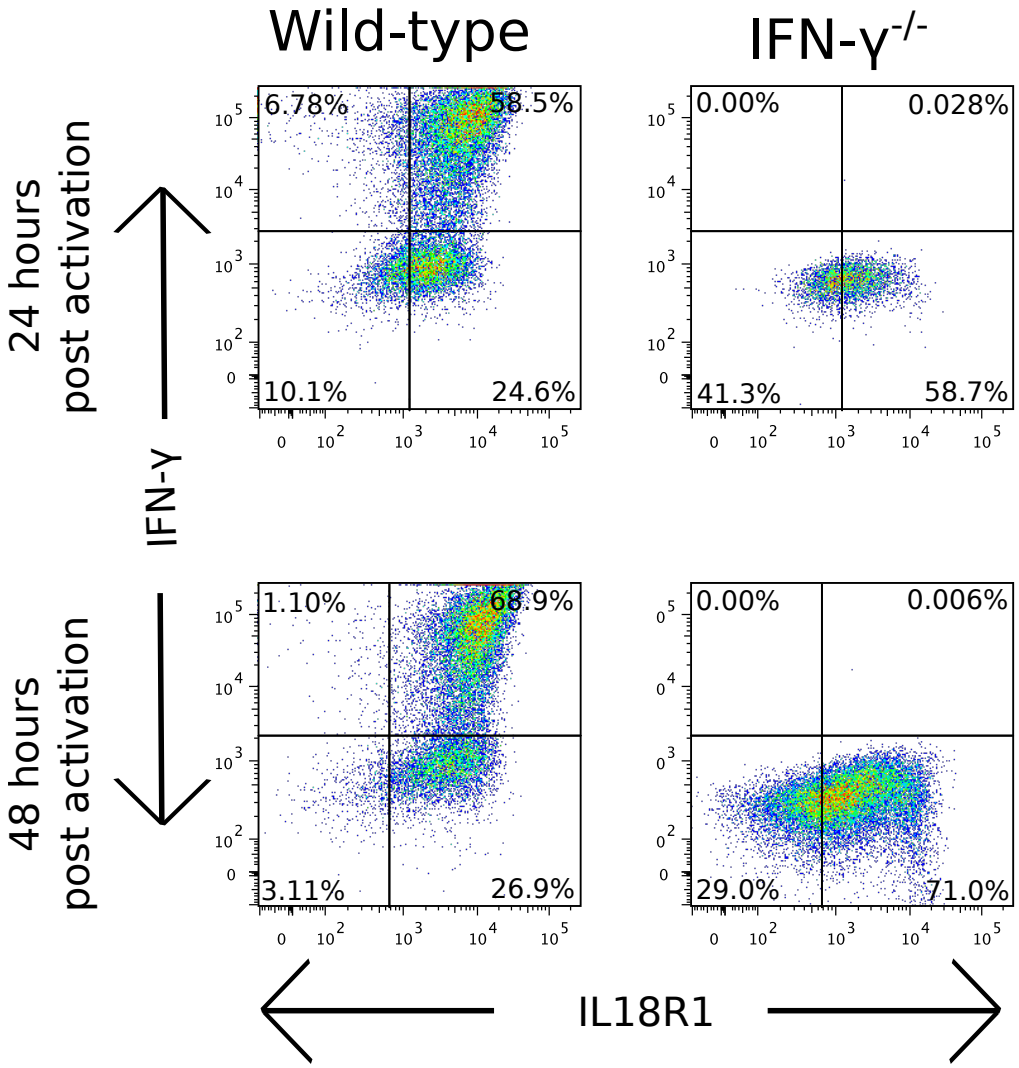

Figure 5.4: IFN-γ deficient cells are able to express IL18R1 in Th1 conditions. - Naïve CD4$^+$ cells were sorted from wild-type and IFN-$\gamma^{-/-}$ mice and activated using αCD3 and αCD28 in Th1 conditions. IL18R1 expression was assessed by flow cytometry at varying timepoints after removal from activation.

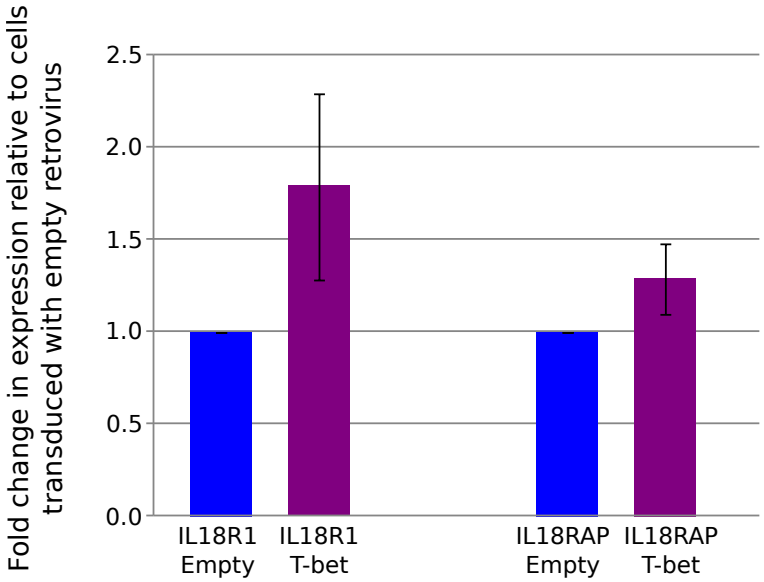

Figure 5.5: T-bet can upregulate both IL18R1 and IL18RAP independently of IFN-γ. - Data from expression arrays on cells from T-bet$^{-/-}$, IFN-$\gamma^{-/-}$ double knockout mice which had been retrovirally transduced with empty retrovirus (Empty) or with constitutively active T-bet (T-bet) was analysed to calculate fold change in expression of IL18R1 and IL18RAP upon addition of constitutively active T-bet. Fold change is normalised to cells transduced with empty retrovirus for each of IL18R1 and IL18RAP. Error bars show standard deviation of two biological array replicates.

had been cultured with just IL-2 for seven days. By contrast, C57BL/6 Thp cells had been sorted as naïve cells directly into TRIzol for RNA extraction. Fold change of mRNA for IL18R1 and IL18RAP across subsets and in wild-type and T-bet$^{-/-}$ cells was calculated relative to mRNA in wild-type Th0 cells (for BALB/c) or wild-type Thp cells (for C57BL/6) (fig. 5.6). Because the neutral (Th0 or Thp) condition is different between each dataset and because datasets were acquired on different types of array, we cannot draw direct comparisons between the two datasets. However, we can examine trends within the datasets for the two strains.

As expected, IL18R1 expression is upregulated in Th1 cells compared to Th0/Thp and although upregulation is seen in the T-bet$^{-/-}$ Th1 cells, this upregulation is less in the T-bet deficient cells than wild-type. IL18RAP expression follows a similar pattern. It is higher in Th1 cells than in Th0/Thp and an increase is still seen but is smaller in T-bet$^{-/-}$ Th1 versus Th0/Thp than in wild-type Th1 versus Th0/Thp. In Th2 cells IL18R1 expression is lower than in Th0/Thp cells but IL18RAP is upregulated both in wild-type and T-bet$^{-/-}$ cells. This suggests that IL18R1 and IL18RAP are not always co-regulated. IL18RAP expression is also upregulated in cells skewed to Th17, whereas IL18R1 expression is either not upregulated (C57BL/6) or very slightly upregulated (BALB/c).

Taken together, these data indicate that IL18R1 and IL18RAP might be co-regulated in Th1 lineage conditions but not in other conditions. The one potential caveat to this is that if IL18R1 is downregulated by CD3, as has been suggested,[190,204–206] stimulation of these cells prior to RNA extraction for the arrays may have affected the results. However, since we do see high levels of IL18R1 in wild-type cells, this does not seem to have been the case.

The two strains of mice that we examined, Balb/c and C57BL/6 are genetically different. To check whether there was any genetic variation between the two strains at either the *Il18r1/Il8rap* locus or the *Il18* locus, we used the Mouse Genome Informatics Resource (http://www.informatics.jax.org/) to search for SNPs on chromosome one (the chromosome on which *Il18r1* and *Il18rap* are located) and chromosome nine (the chromosome on which *Il18* is located). There were no SNPs reported between C57BL/6 and Balb/c in any of the loci examined. The nearest SNP to the *Il18r1/Il8rap* locus was over 470kbp away and the

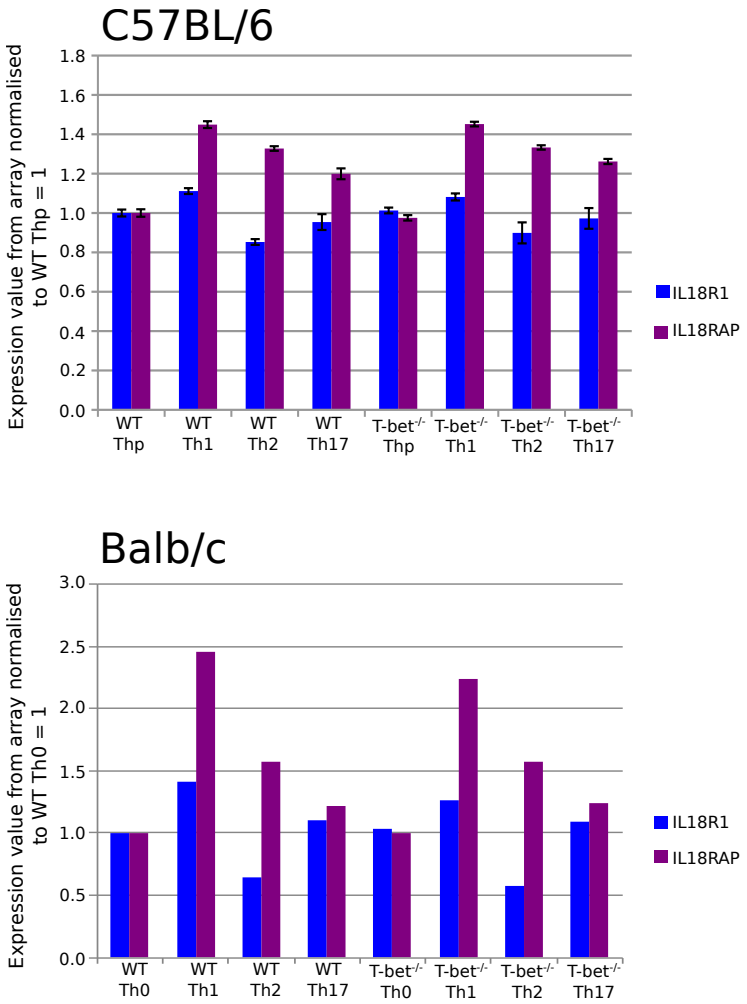

Figure 5.6: Expression of IL18R1 and IL18RAP across CD4$^+$ cell subsets. - Data from expression arrays on different CD4$^+$ cell subsets in both BALB/c and C57BL/6 mice was analysed for expression of IL18R1 and IL18RAP. Data was normalised to expression for wild-type (WT) cells in Th0 or Thp condition. Data for C57BL/6 is averaged over two replicate arrays. Data for Balb/c is from one array only.

nearest SNP to the *Il18* locus was over 170kbp away.

5.3 IL18R1 Expression *In Vivo*

The above work demonstrates some key principles of IL18R1 regulation but in optimized *in vitro* culture conditions where populations can be neatly defined and tightly controlled. As discussed in the introduction and elsewhere, while the $CD4^+$ lineage subset paradigm is a useful concept and broadly applicable, the *in vivo* context is not always quite as well delineated or as fixed. To investigate the role of IL18R1 *in vivo* in a disease context, we used two mouse models. Because our interest in the *IL18R1* locus stemmed from an observation of differential T-bet binding at a SNP in a region highlighted by GWAS for coeliac and Crohn's diseases, we used a model of general Inflammatory Bowel Disease (IBD) and a model of coeliac disease.

5.3.1 Naïve T Cell Transfer Model of Inflammatory Bowel Disease

For the IBD model, we sorted naïve cells from wild-type or T-bet $^{-/-}$ mice and injected 500,000 of these cells per mouse into $Rag2^{-/-}$ hosts. We sacrificed half of the recipient mice at two weeks after injection and the remaining mice at four weeks after injection. We harvested spleen and mesenteric lymph nodes to examine kinetics of IL18R1 upregulation across disease timepoints in addition to across genotypes. We stimulated the cells with PMA and Ionomycin and assayed for IL18R1 and IFN-γ by intracellular staining and flow cytometry (fig. 5.7). Despite low cell counts in some conditions, this model shows that T-$bet^{-/-}$ cells can express IL18R1 during an *in vivo* model, as well as in optimised *in vitro* culture. Interestingly, in the T-$bet^{-/-}$ cells we saw a decrease in IL18R1 expression from two weeks to four weeks suggesting that IL18R1 expression is less stable with time in the absence of T-bet which agrees with our *in vitro* findings. This experiment needs repeating to confirm these results.

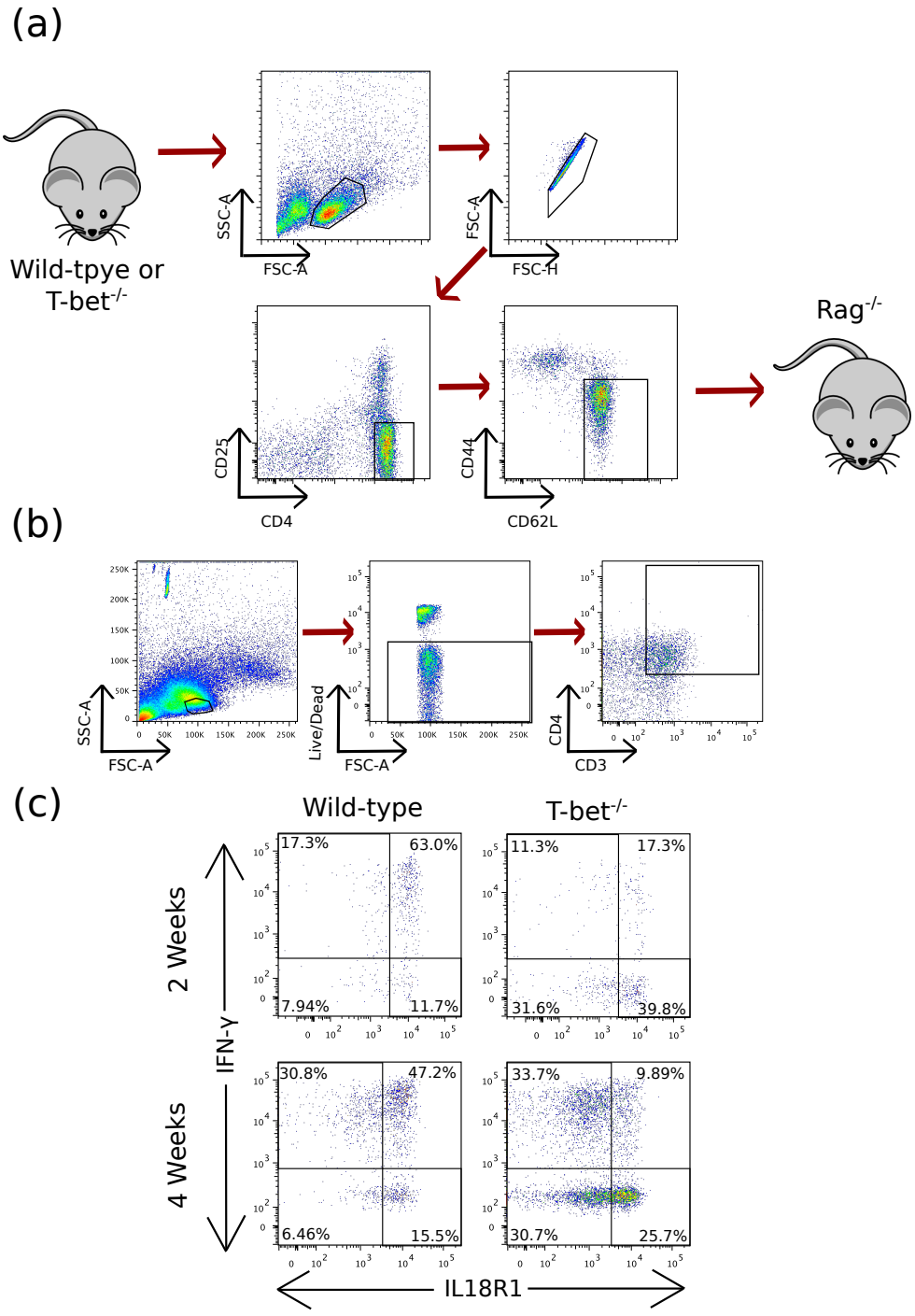

Figure 5.7: IL18R1 expression with progression of naïve adoptive transfer model of IBD. - Rag2$^{-/-}$ mice were injected with naïve CD4^{+} cells as shown in (a). Cells were recovered at two weeks or four weeks and CD4^{+} cells were examined for expression of IFN-γ and IL18R1 as shown in (b). Results are shown in (c).

5.3.2 Primed Effector/Memory Cell Model of Coeliac Disease

For the coeliac model, we used the model published by Freitag *et al*.[327] We immunized wild-type and T-bet$^{-/-}$ mice with gliadin in adjuvant and then transferred the effector/memory CD4^{+} cells from these mice into Rag1 deficient hosts (fig 5.8). If maintained on a diet containing high levels of gluten these mice develop symptoms resembling coeliac type disease. Similar to its human counterpart, the model is milder than most models of IBD. However, administration of wild-type effector/memory cells to Rag1$^{-/-}$ recipients has been shown to slow subsequent weight gain by the mice compared to controls and to result in inflammation and increased IFN-γ, and possibly IL-17, expression in the duodenum.

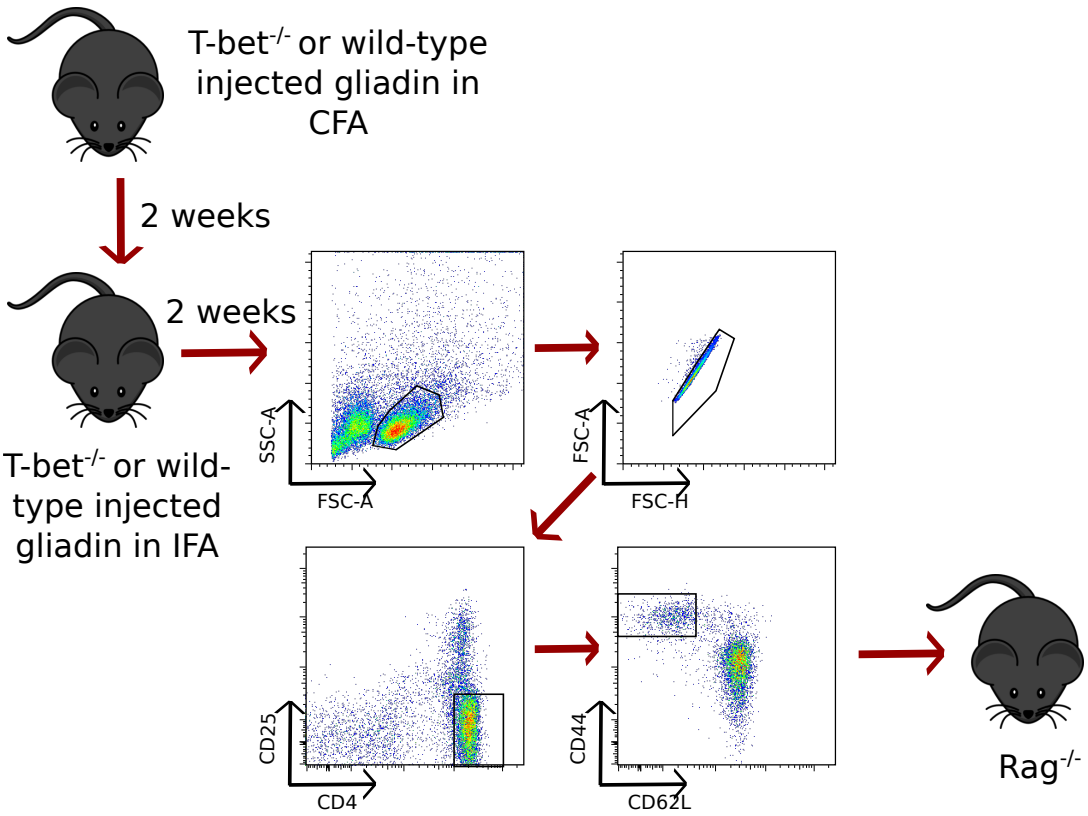

Figure 5.8: Coeliac disease model. - Schematic of coeliac disease model used.

In agreement with Frietag *et al*, we found that, although all mice gained weight during the eight week timecourse, mice given effector/memory CD4^{+} cells did tend to gain less weight as a percentage of starting body mass than mice given PBS control (fig. 5.9). However, in our hands, this was not significant.

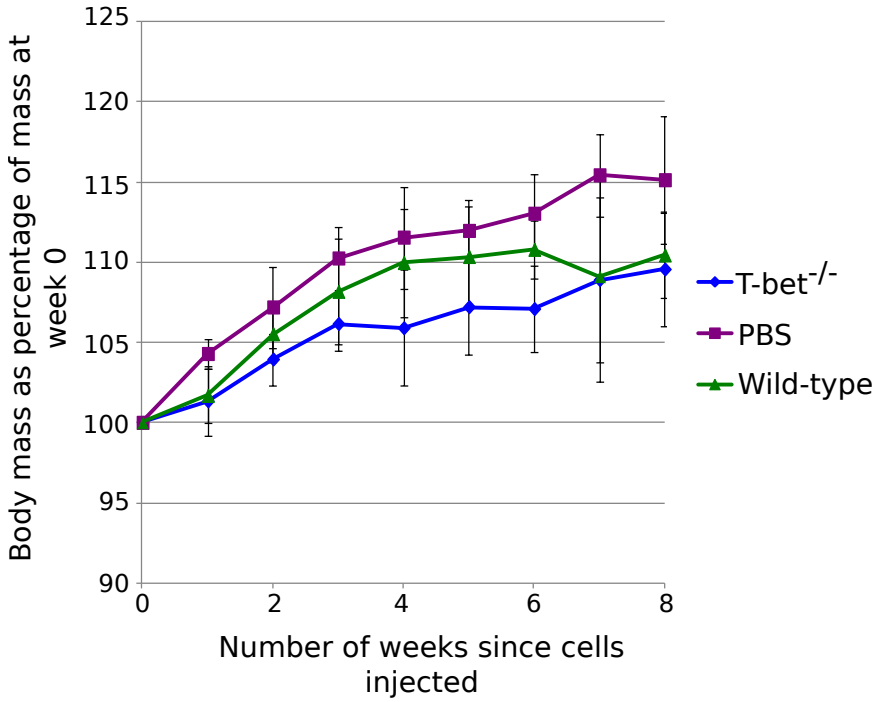

Figure 5.9: Weight gain in coeliac model. - Mice were injected with effector/memory $CD4^+$ cells from wild-type or T-$bet^{-/-}$ mice that had been primed with gliadin injection or with PBS as control. Weight was measured over eight weeks and is shown relative to mouse weight at time of injection. Error bars show standard deviation in percentage weight change relative to starting weight for all mice in group. Group sizes were as follows: three mice received wild-type effector/memory cell, four mice received T-$bet^{-/-}$ effector/memory cells, two mice received PBS.

After eight weeks, we sacrificed the mice. We weighed the spleen and small intestine of each mouse (fig. 5.10). We saw no correlation between mass of either organ and receipt of wild-type or T-bet$^{-/-}$ cells or PBS control.

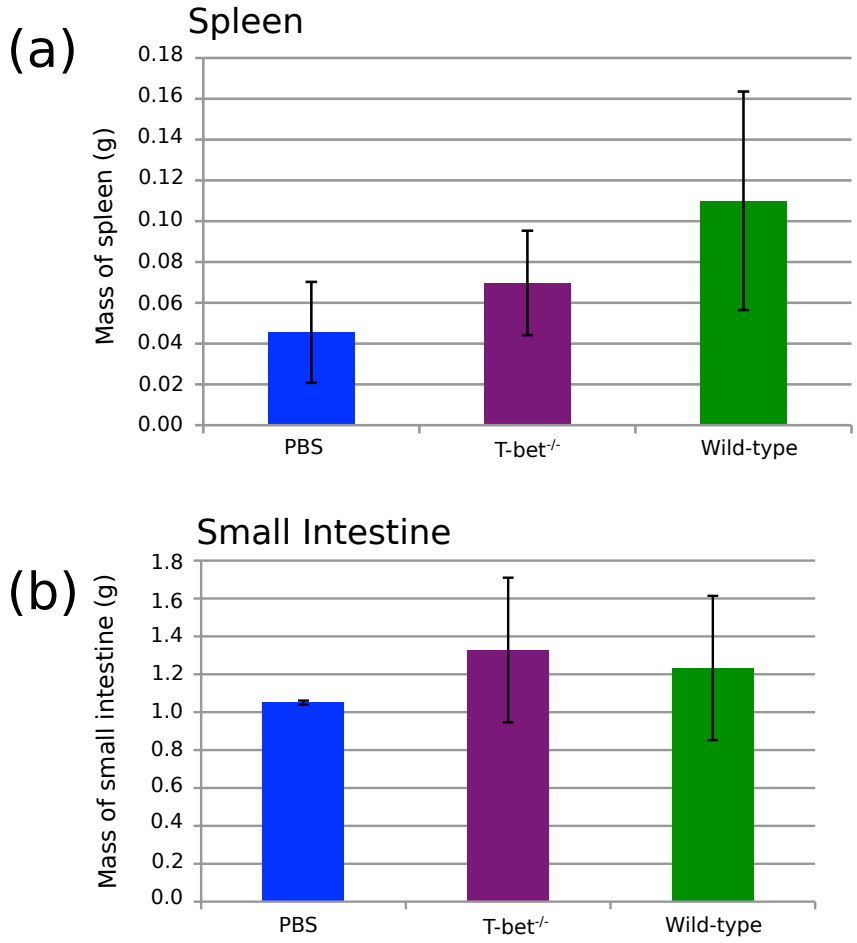

Figure 5.10: Mass of spleen and small intestine showed no trend between experimental and control animals. - Spleen mass is shown in (a). Mass of small intestine is shown in (b). Error bars represent standard deviation across experimental group.

We stained a section of the proximal small intestine for histology. This showed some disease in some mice. We saw partial villous atrophy in some sections (fig. 5.11 (a) and (b)) and fat villi and eosinophil infiltrate were also seen in some mice. However, crypt hyperplasia (as measured by number of mitotic cells in five adjacent crypts) and inflammation (as scored by an external assessor from 0 to 3 for Lamina Propria Chronic Inflammation (LP CI)) did not correlate with receipt of wild-type or T-bet$^{-/-}$ cells or PBS control (fig 5.12). As previously

mentioned, the model is relatively mild. Furthermore, the mice are monitored over an eight week period after receipt of cells. This constituted sufficient time for one of our PBS controls to develop inflammation despite having not received cells, possibly from reaction to a microbe not screened out by our in-house Specific Pathogen Free conditions and in the context of an immunodeficient state. This mouse had a small spleen, partial villus atrophy and scored the maximum LP CI score of three. These findings strongly suggests that we had too few mice to draw strong conclusions about overall disease at this time. Of note, Freitag *et al* also find duodenal inflammation in their control group, although it is not as great as in their experimental groups.

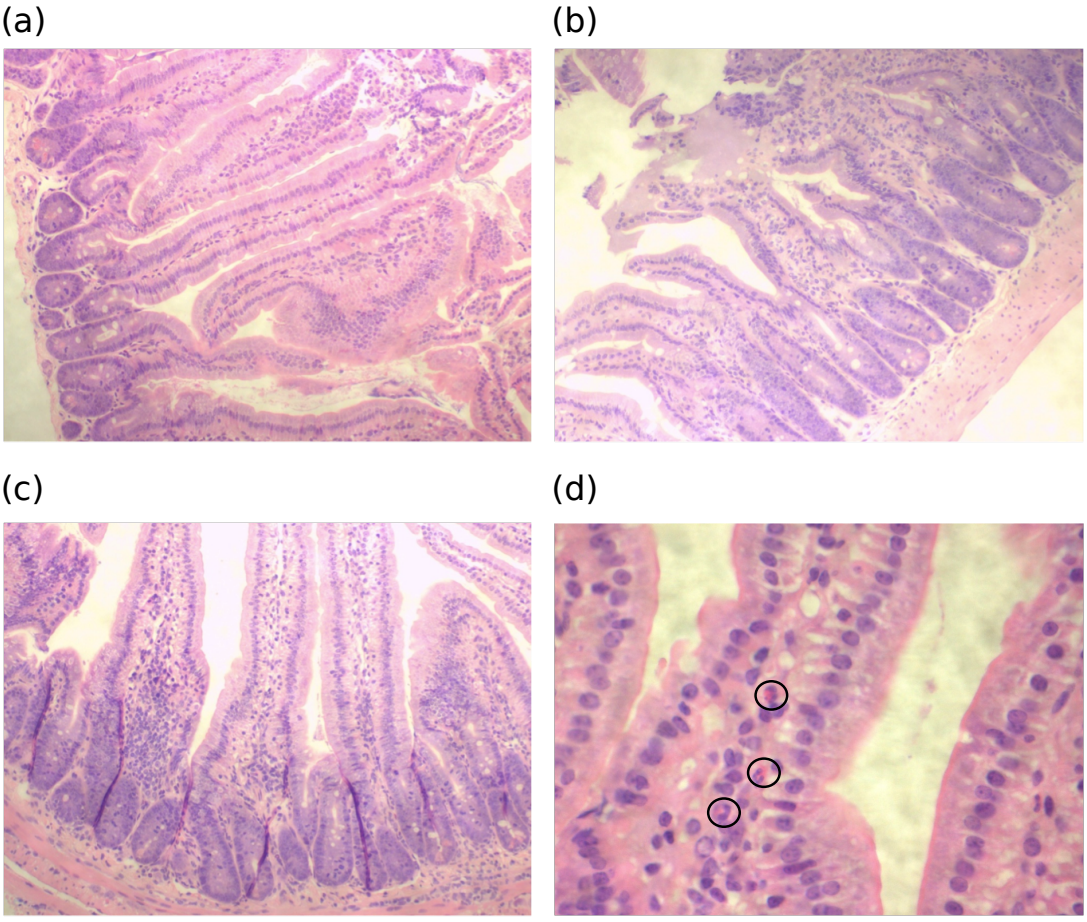

Figure 5.11: Histology of small intestine sections showed some evidence of disease in coeliac model. - Sections of small intestine were harvested and stained for each mouse in coeliac model. (a) Normal small intestine from PBS control mouse. (b) Example of partial villous atrophy in small intestine of mouse injected with T-bet$^{-/-}$ cells. (c) Example of fat villi in mouse injected with wild-type cells. (d) Example of eosinophil infiltration in mouse injected with wild-type cells. Eosinophils circled in black.

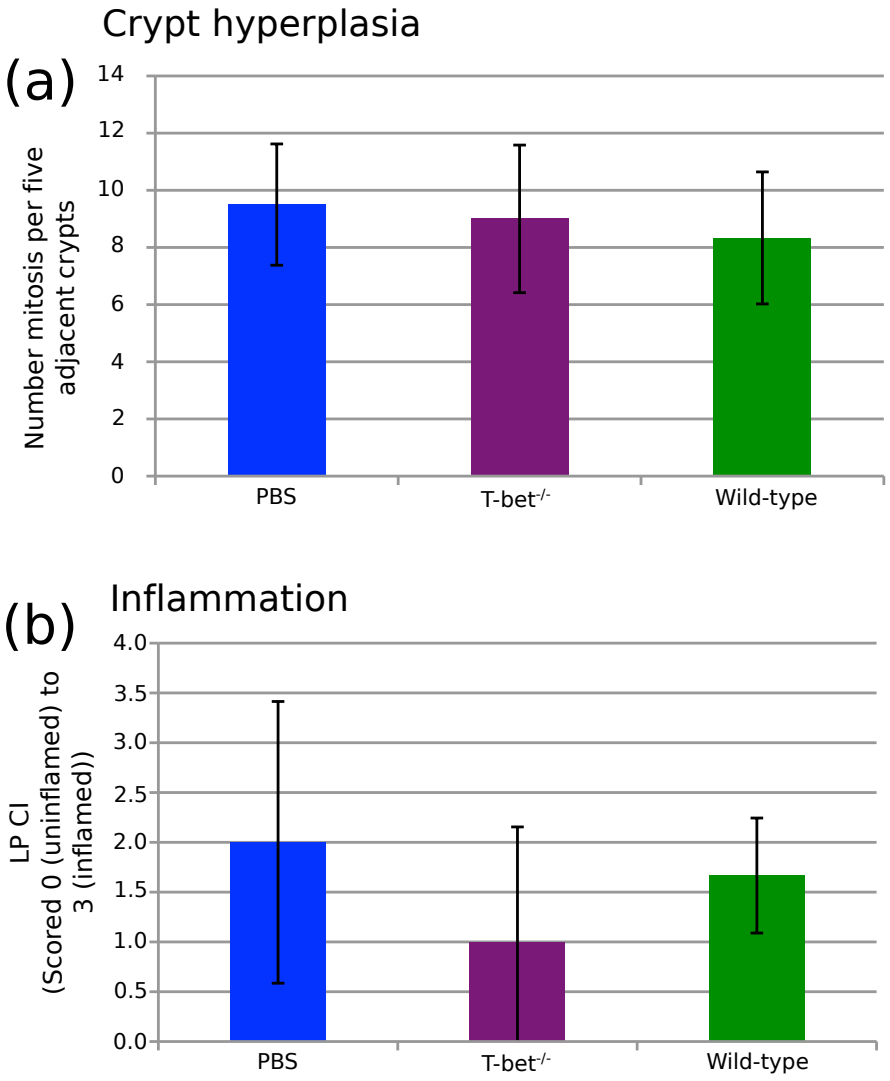

Figure 5.12: Histology scores for coeliac model. - Sections of small intestine were examined for presence of mitotic cells (a) and scored for presence of inflammation (b). Error bars indicated standard deviation for experimental group.

To examine the cytokine production of the cells injected, we harvested the mesenteric lymph nodes and analysed IL18R1 and IFN-γ expression on the $CD4^+$ cells found in both spleen and mesenteric lymph node by intracellular staining. We also examined the expression of IL-17A on these cells because the research by Frietag *et al*[327] suggests that IL-17 has a role in the disease in addition to IFN-γ. Unsurprisingly, we found fewer IFN-γ producing cells in the T-bet$^{-/-}$ samples. We found a high number of IL-17A producing cells and this number was further increased in the T-bet$^{-/-}$ samples (fig. 5.13). In the mesenteric lymph node, the difference in percentage of IFN-γ producing cells between the two groups was significant (p = 0.02).The differences in the other populations was not. In the spleen, the difference in percentage of IFN-γ producing cells was also significant (p = 0.01), as was the difference in double producers (IFN-γ^+, IL-17A$^+$ (p= 0.03). On closer examination of the plots (fig. 5.13 (b)) we found that one mouse from the T-bet$^{-/-}$ group had very low numbers of IL-17A producing cells compared to all others in the group. We checked for any other abnormalities in this mouse and found that it also had a lower small intestine mass than all other mice receiving T-bet$^{-/-}$ cells. Furthermore, we recovered fewer $CD4^+$ cells from this mouse than the others suggesting that the model may not have worked properly in this mouse. If we exclude this sample from analysis then, in the lymph node, the percentage of IFN-γ producing cells remains significant (p=0.01) and the percentage of IL-17A producing cells become significant (p= 0.01). In the spleen, percentage of IFN-γ producing cells also remains significant (p= 0.026) and the percentage of IL-17A producing cells also becomes significant (p= 0.008). However, when this sample is removed the percentage of double producing cells is no longer significant (p=0.059).

We then examined IL18R1 expression in the $CD4^+$ population. Most of the $CD4^+$ population were IL18R1 positive. (fig. 5.14 (a)). To show that this was not due to unspecific binding, we compared the $CD4^+$ population to the total live cell population. In the total live cells, we saw a bimodal distribution of IL18R1 expression as we would expect from the different cells types present and demonstrating that out results did not result from non specific binding (fig. 5.14 (b)).

Although most $CD4^+$ cells were IL18R1 positive, we did see a few IL18R1 negative cells (as seen from the negative 'tails' on the histograms in fig. 5.14 (a)). To investigate whether these belonged to a specific population of cytokine pro-

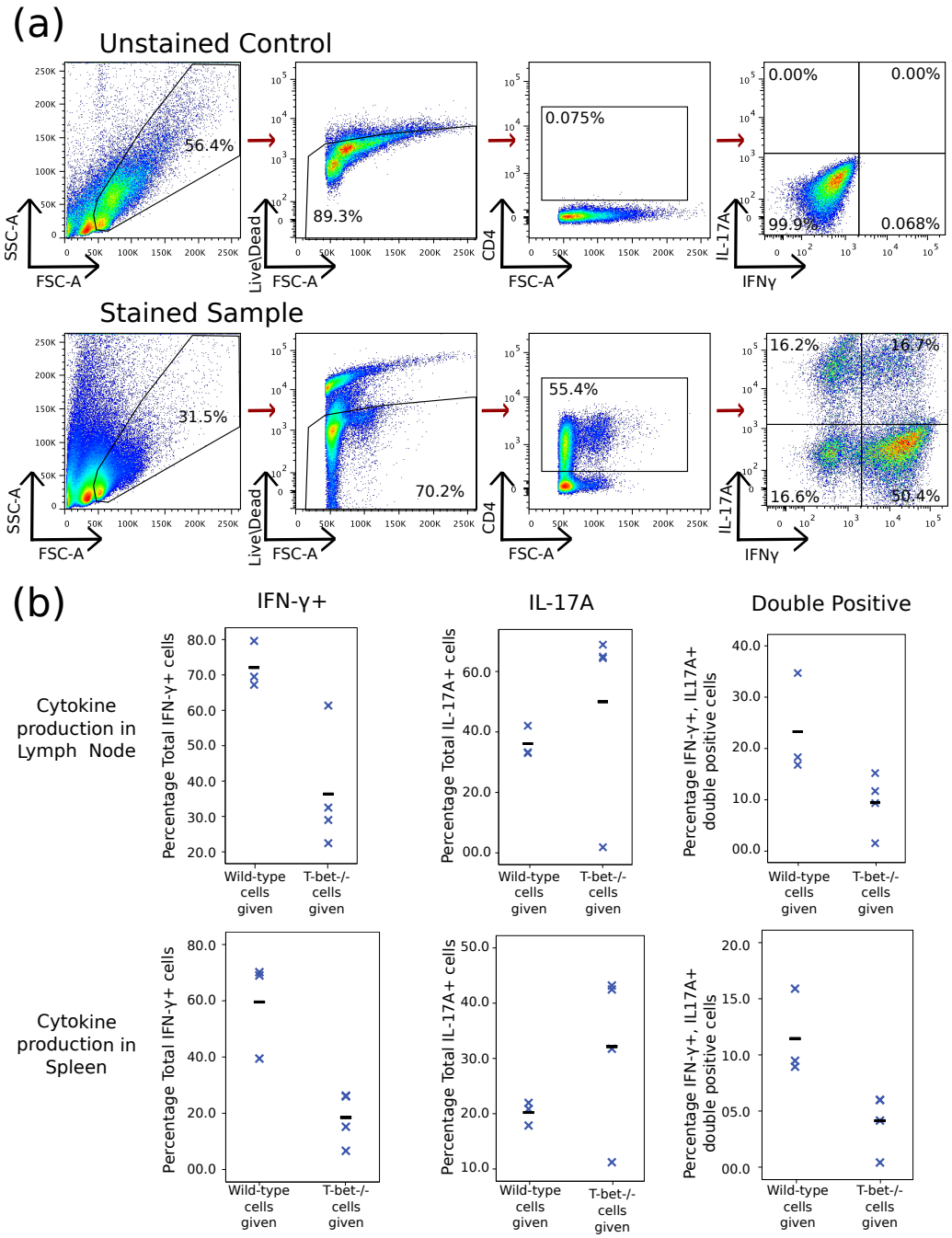

Figure 5.13: Cytokine production in spleen and mesenteric lymph node from coeliac model. - Spleen and mesenteric lymph nodes were harvested eight weeks after administration of gliadin primed effector/memory $CD4^+$ cells and stained to asses IL-17A and IFN-γ production by intracellular stain. Gating protocol for experiment is shown in (a). Percentage of IL-17A producers, IFN-γ producers and IL-17A, IFN-γ double producers in mice injected with wild-type or T-bet$^{-/-}$ cells is shown in (b). Blue crosses represent individual mice and black lines represent mean for group.

ducers, we examined IL18R1 expression in the different cytokine producing populations of the $CD4^+$ population. (($IFN\text{-}\gamma^+$, $IL\text{-}17A^-$), ($IFN\text{-}\gamma^+$, $IL\text{-}17A^+$), ($IFN\text{-}\gamma^-$, $IL\text{-}17A^+$) and ($IFN\text{-}\gamma^-$, $IL\text{-}17A^-$)) (fig 5.15). As can be seen from the figure, the IL-17A producing populations (red and blue lines) tended to have high levels of IL18R1 expression independently of genotype. By contrast the IL-17A negative populations and, particularly the $IFN\text{-}\gamma^+$ $IL\text{-}17A^-$ population showed a bimodal distribution of IL18R1 expression in the $T\text{-}bet^{-/-}$ cells. This second result demonstrates the need for T-bet to stabilise IL18R1 expression in Th1 type cells. However, the T-bet independent high expression of IL18R1 on IL-17A single producers and $IL\text{-}17A^+$, $IFN\text{-}\gamma^+$ double producers is more unexpected and suggests a role for IL18R1 in non Th1 type cells. This result also suggests IL18R1 expression is T-bet independent in non Th1 type cells in this model.

Gating for the above results was based on unstained control (see figure 5.13). However, some $CD4^+$ cells were seen in mice injected with PBS control. This was most likely because $CD4^+$ is also expressed at low levels on some macrophages. To exclude the possibility that low level expression of CD4 on macrophage were affecting our results, we redid the $CD4^+$ gating based on a PBS control mouse rather than the unstained control. This affected some of our percentages but did not affect the overall trends in our data. (Data not shown.)

We also harvested a small section of the small intestine from each mouse, extracted the RNA and examined expression of IL18R1, IL18RAP, IFN-γ and IL-17A by qPCR. The results showed a general trend towards more cytokine production in those mice that had received cells versus PBS control and this was significant in IFN-γ production between mice given wild-type cells and PBS ($p=0.017$). This suggests that the injected $CD4^+$ cells were instigating active inflammation in the small bowel despite the mixed results from the histology data. However, no significant differences were found between mice given wild-type and mice given $T\text{-}bet^{-/-}$ cells, although IL18R1 expression tended towards significance ($p=0.089$) (see fig. 5.16).

5.3.3 Cells used in Coeliac Model

We also phenotyped the effector/memory $CD4^+$ cells that are generated in the wild-type or $T\text{-}bet^{-/-}$ mice by gliadin priming. We injected wild-type and T-

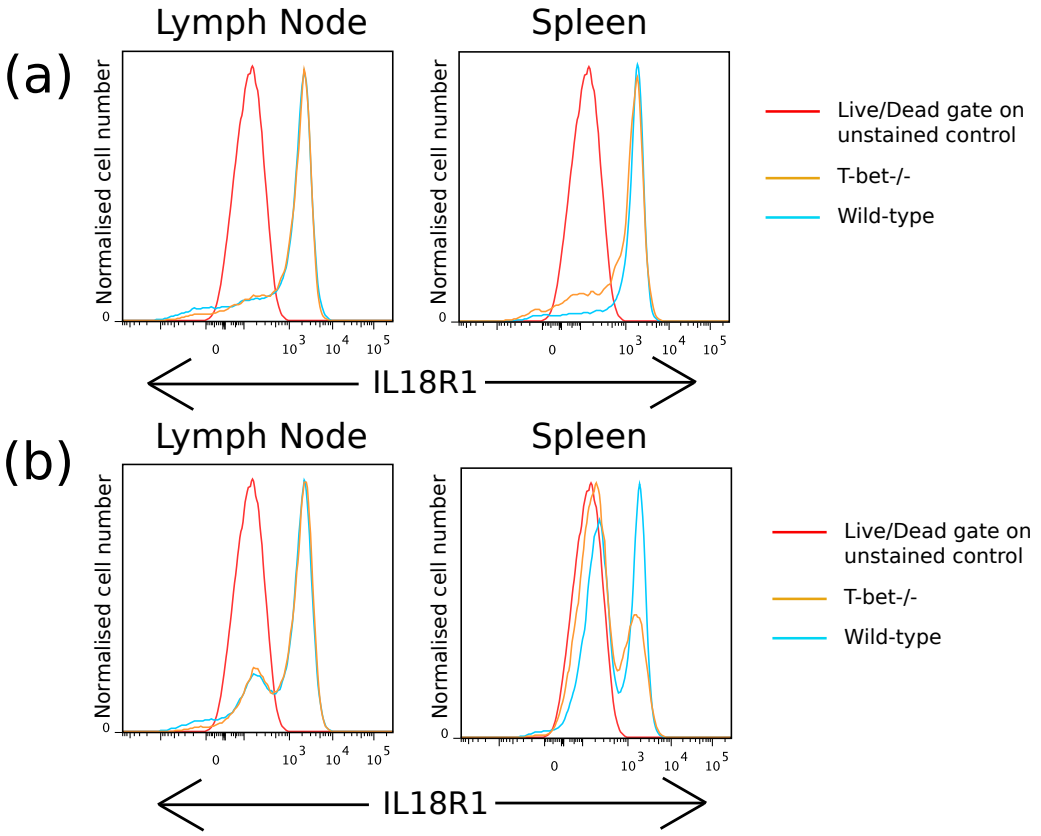

Figure 5.14: Most CD4$^+$ cells from coeliac mice are IL18R1 positive. - IL18R1 expression was examined in CD4$^+$ cells and total live cells from coeliac model mice. (a) Example plots for IL18R1 expression in CD4$^+$ cells from spleen and mesenteric lymph node of mice injected with wild-type cells (blue line) or T-bet$^{-/-}$ cells (orange line). (b) Example plots for IL18R1 expression in total live cells from spleen and mesenteric lymph node of mice injected with wild-type cells (blue line) or T-bet$^{-/-}$ cells (orange line).

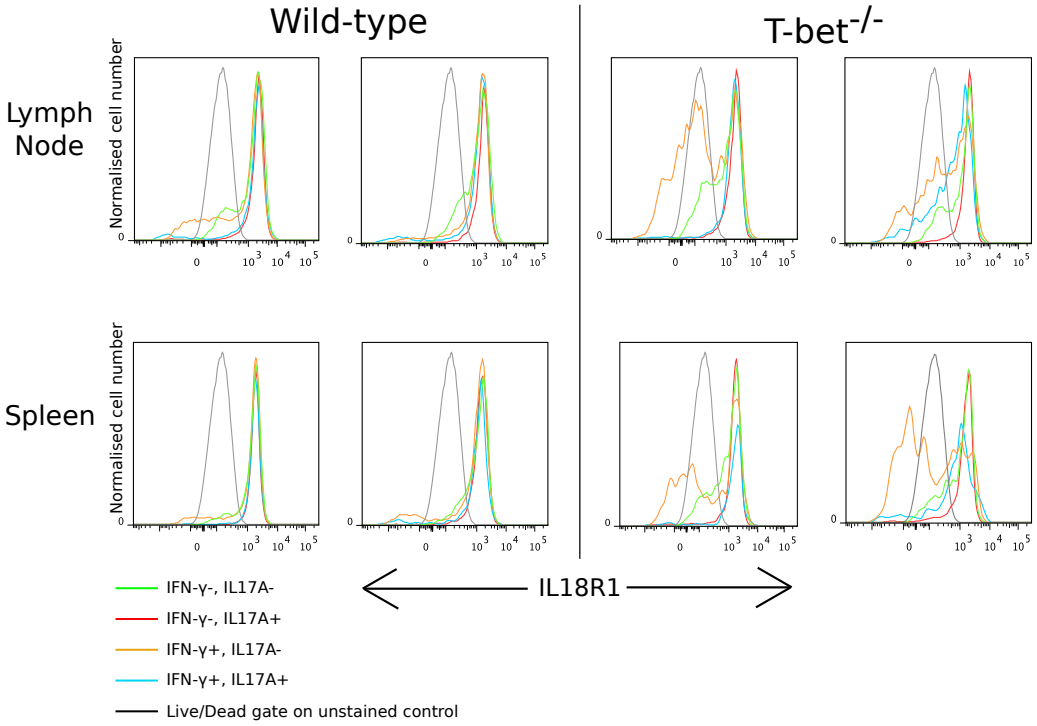

Figure 5.15: Example plots for IL18R1 expression on subsets of cells taken from spleen and lymph node of mice in coeliac model. - All cells are $CD4^+$ and then further gated on IFN-γ^-, $IL17A^+$ (red line), IFN-γ^-, $IL17^-$ (green line), IFN-γ^+, $IL17A^-$ (orange line) or IFN-γ^+, $IL17^+$ (blue line). Two plots are shown per condition and each plot represents one mouse. All histograms show cell count normalised to total cells in the gate.

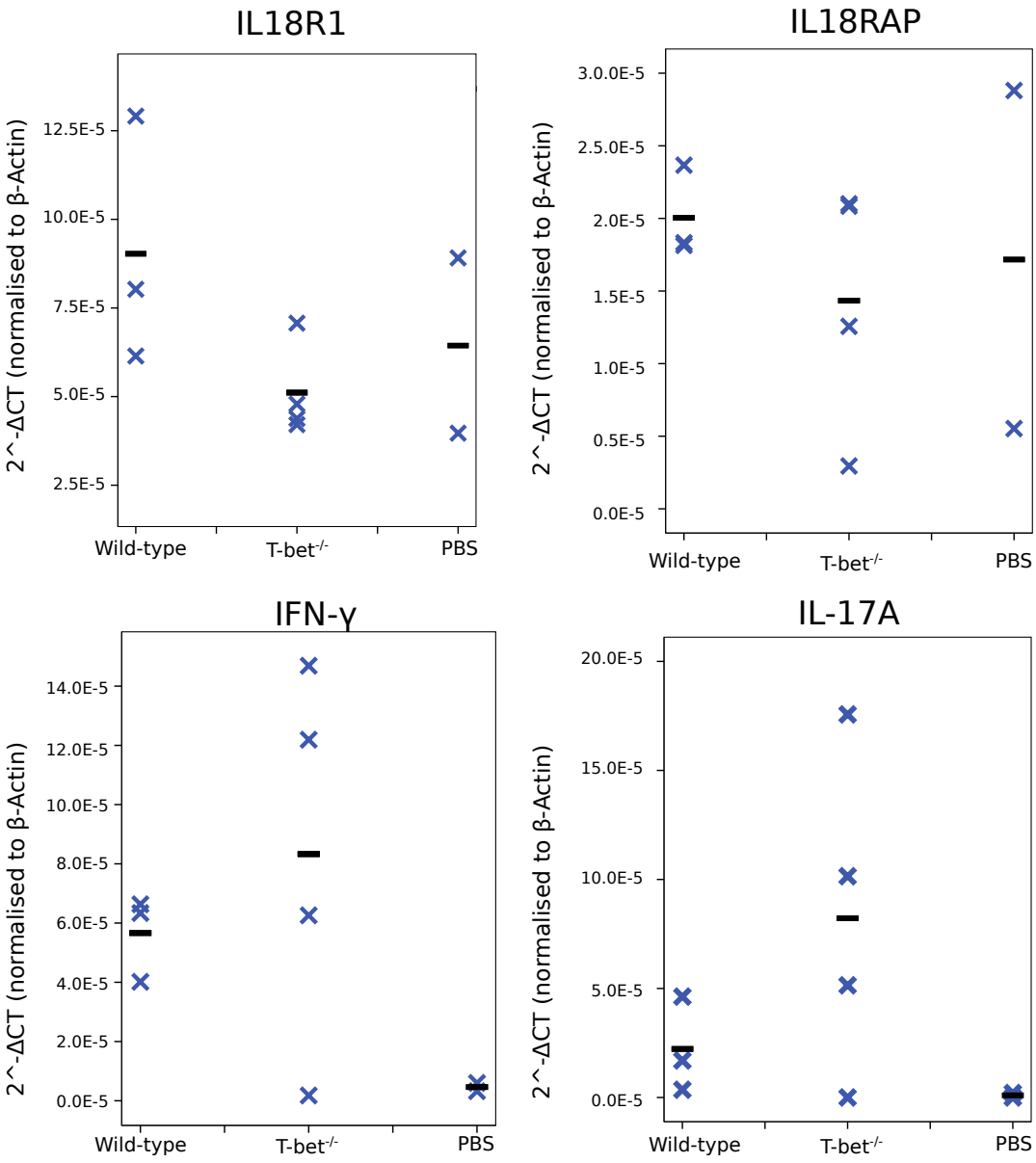

Figure 5.16: MessengerRNA expression in small bowel of coeliac mice. - Tissue was harvested from small bowel of mice in coeliac model and extracted RNA was assayed for expression of IL18R1, IL18RAP, IFN-γ and IL-17A

bet$^{-/-}$ mice with gliadin in Complete Freund's Adjuvant (CFA) and then Incomplete Freund's Adjuvant (IFA) as before or with CFA and then IFA on its own or with two rounds of PBS. We then harvested spleens as before. However, instead of sorting effector CD4^{+} cells and reinjecting into Rag1$^{-/-}$ hosts, we stimulated the cells and examined expression of IL18R1, IFN-γ and IL-17A. We also stained for most of the markers we had used for sorting the cells (CD4, CD44, and CD25) to test for changes in the number of naïve and effector/memory CD4^{+} cells in the mice as a result of gliadin priming. Surprisingly this suggested that the mice primed with PBS had a greater percentage of effector and smaller percentage of naïve cells as a percentage of total CD4^{+} than mice given gliadin in adjuvant or adjuvant alone (fig 5.17).

However, CD62L is downregulated by activation with PMA and Ionomycin and so gating for effector/memory cells was on CD44 expression alone. Furthermore, CD25 is upregulated by activation with PMA and Ionomycin so we could not use this as a marker to exclude Tregs from our analysis as excluding CD25^{+} cells also excluded many effector cells. Since Treg cells comprise approximately 5-10% of total CD4^{+} cells, this was not ideal.[328] To address some of these issues, we examined cells from some of the mice for surface markers only, as used for cell sorting, without stimulation. This showed that, as expected, the mice injected with PBS had fewer effector/memory and more naïve cells than those mice injected with gliadin in CFA/IFA or with CFA/IFA alone. In the wild-type mice, we saw more effector cells in those mice given gliadin in addition to CFA versus those mice given CFA alone. In the T-bet$^{-/-}$ mice, we saw a higher percentage of cells in those mice just given CFA (fig. 5.18). However, for the surface staining, we only have two mice per group so cannot draw firm conclusions.

Unsurprisingly, we saw very little cytokine production in our naïve cells (see fig 5.19 (a) for example data). In the effector/memory population, we found very few IFN-γ producing cells from the T-bet$^{-/-}$ mice, whereas between 10% and 40% of cells from wild-type mice produced IFN-γ. IL-17A producing cells were seen in the effector/memory population from all mice but were generally slightly more numerous from the T-bet$^{-/-}$ mice. These results might be expected from the role of T-bet in IFN-γ production. Less expected was a similar trend to the effector/memory cell counts in that mice injected with PBS tended to have more cytokine producing cells than mice from the same genotype injected with

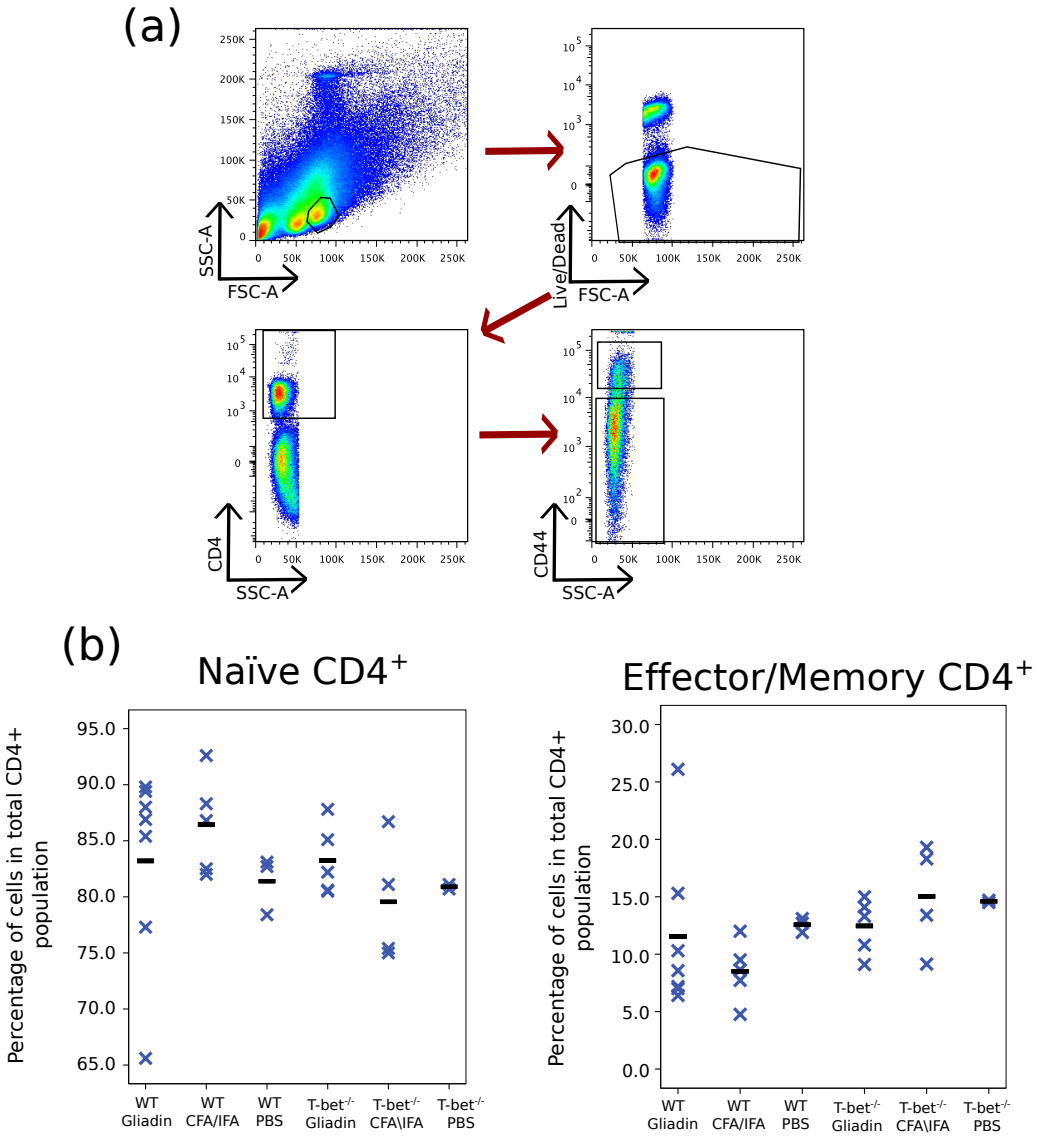

Figure 5.17: Phenotype of cells introduced into coeliac model. - Wild-type or T-bet$^{-/-}$ mice were injected with gliadin, CFA/IFA alone or PBS and percentage of naïve and effector memory CD4^{+} cells was measured using CD44 expression. Gating for populations is shown in (a). Percentage of naïve and effector memory as percentage of total CD4^{+} is shown in (b). Blue crosses represent individual mice. Black lines represent mean for the group.

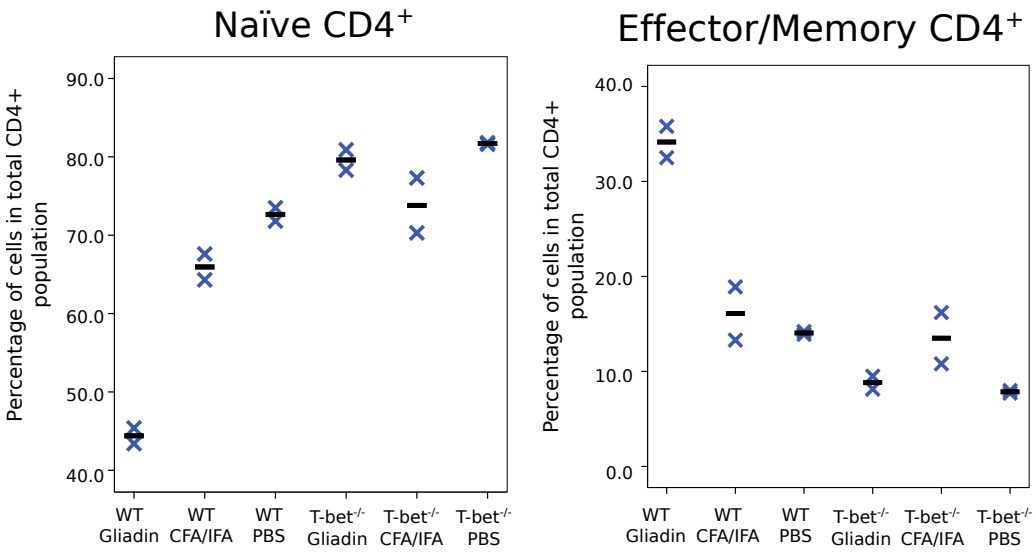

Figure 5.18: Phenotype of cells introduced into coeliac model based on CD44 and CD62L expression in unstimulated population. - Cells were harvested as before but surface stained only without stimulation based on surface expression of CD44 and CD62L.

gliadin in CFA/IFA or CFA/IFA alone (fig 5.19). However this may be due to the surface marker gating issues discussed earlier.

We found very few double positive (IFN-γ^+, IL-17A$^+$) cells in the effector/memory fraction of our cells. (For one wild-type PBS treated mouse we found 3.08% and, for another, 2.33% but all others were under 2%.). However, we did find cells that stained positive for IL18R1 and also expressed IFN-γ which is somewhat expected and cells that were positive for IL18R1 and also expressed IL-17A which was less expected, although consistent with the results from the full coeliac model. Percentages of the effector/memory population that were double positive for IL18R1 and IFN-γ or double positive for IL18R1 and IL-17A are shown in figure 5.20.

The surface stain experiment also allowed us to examine the difference in IL18R1 expression between effector/memory and naïve CD4$^+$ cells. Our analysis supported previous reports that, in naïve T cells, there is a basal level of IL18R1 but this is either up- or downregulated on activation depending on local conditions to give a bimodal distribution in effector/memory cells (fig. 5.21). We have yet to determine whether IL-18 signalling can occur through this basal level of IL18R1 on naïve CD4$^+$ cells and whether cytokines other than IFN-γ can be produced as a result.

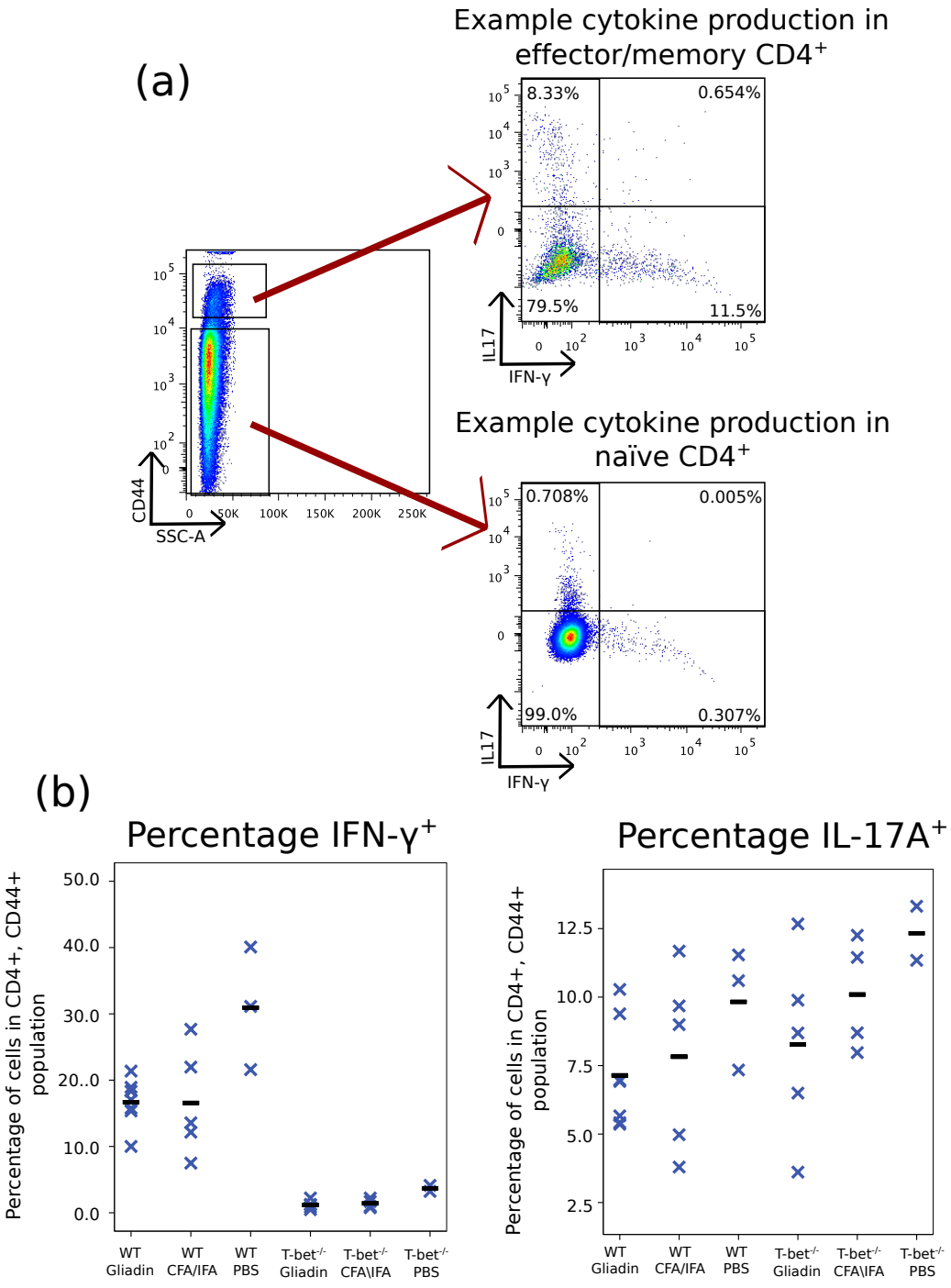

Figure 5.19: Cytokine production by cells used in coeliac model. - Cells from gliadin primed mice were analysed for IFN-γ and IL-17A production. Gating for naïve and effector memory cells and example plots for cytokine production are shown in (a). Percentages of IFN-γ producers and IL-17A producers in effector/memory population is shown in (b).

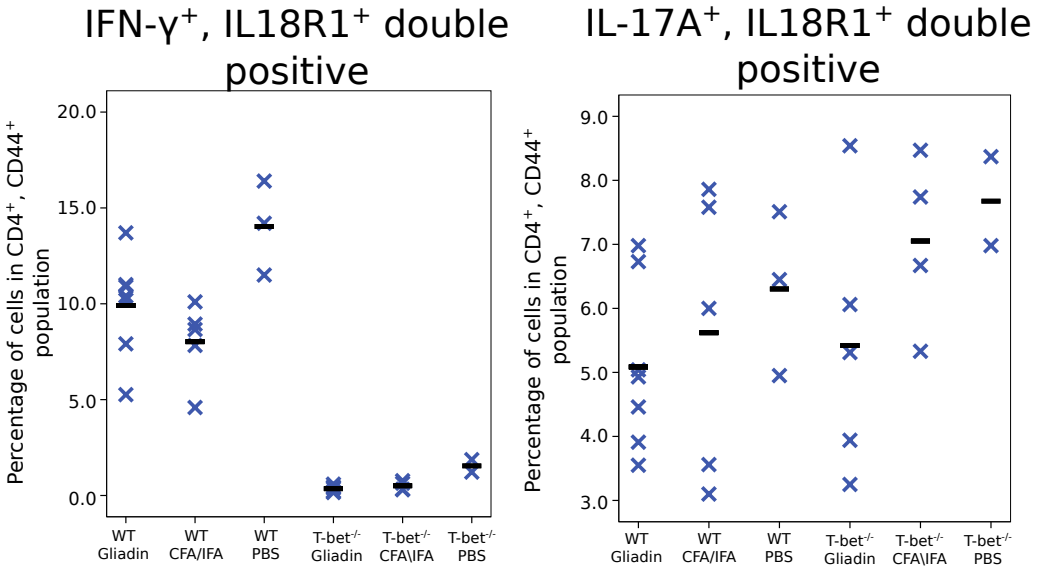

Figure 5.20: IL18R1 expression across IFN-γ and IL-17A producing cells. - Cells were analysed for expression of IL18R1 plus production of either IFN-γ or IL-17A.

5.4 Conclusion

We have confirmed reports of basal IL18R1 expression on unstimulated $CD4^+$ cells. Because we sorted our cells to high purity of naïve cells, we can conclude that IL18R1 is expressed on naïve $CD4^+$ cells. Expression seen in previous papers does not all result from contamination with an effector/memory population. Our results suggest that IL18R1 and IL18RAP are co-regulated in Th1 cells and that T-bet is required for full expression of IL18R1 in these cells. However, some IL18R1 can be expressed in the absence of T-bet and T-$bet^{-/-}$ cells can signal in response to IL-18 *in vitro*.

Our coeliac model shows some interesting data as to cytokine production and IL18R1 expression with respect to T-bet loss *in vivo*. However, the model was generally inconsistent and subject to high experimental variability as seen from our macroscopic and histology data. Results from this model should therefore be treated with caution and we cannot draw any firm conclusions about overall disease pathology at this time. It is possible that larger sample sizes would have allowed us to make firmer conclusions. However, this experiment was a preliminary experiment to examine whether we could establish a coeliac model and whether such a model was worth using. As such, the experimental group sizes

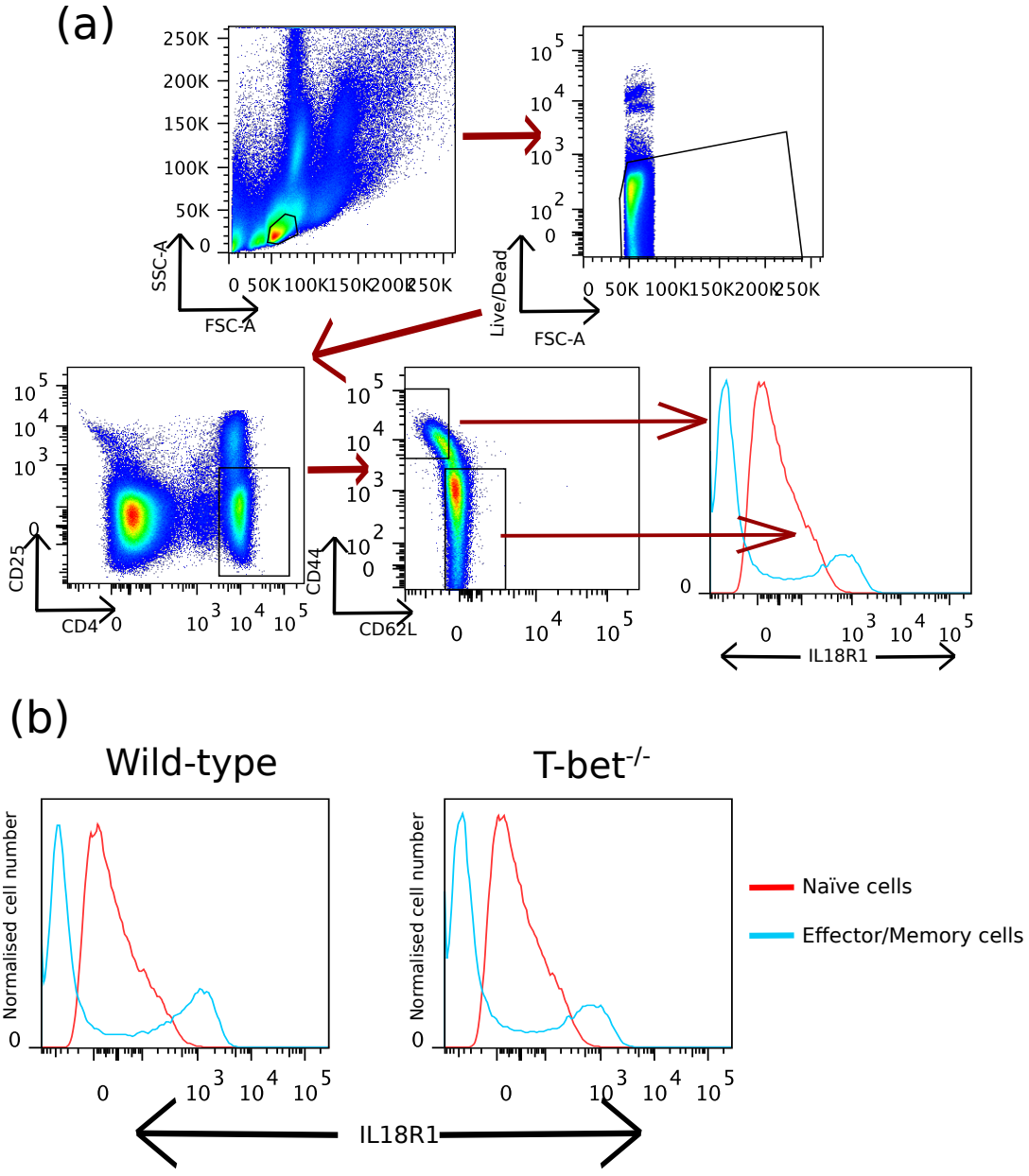

Figure 5.21: Bimodal distribution of IL18R1 seen on effector/memory cells. - IL18R1 expression on cells was analysed on naïve and effector memory populations. Example gating is shown in (a). Example histograms showing IL18R1 expression in naïve (red line) and effector/memory (blue line) populations, for both wild-type and T-bet$^{-/-}$ cells, is shown in (b).

used were small. Given the high variability in the data and the practicalities of needing to use different genotypes of mice at different stages of the experiment, we conclude that this model is not suitable for further use in our work.

However there is some suggestion that, in a disease context, high levels of IL18R1 can be expressed on IFN-γ^- cells both in wild-type and T-bet $^{-/-}$ cells hinting at a distinct T-bet independent mechanism of IL18R1 upregulation in cells that are non Th1 like. However, this requires further investigation.

6

Discussion

6.1 Overview

Recent large scale genomics projects such as ENCODE and the 1000 Genomes Project have demonstrated the abundance of SNPs across the genome and have suggested that SNPs in regulatory regions, including transcription factor binding sites, have the potential to influence downstream phenotype including disease susceptibility. Such work builds on the increasing quantity of data on SNPs associated with diseases and other traits by GWAS. At present, most of the results presented are based on statistical associations. For example, the 1000 Genome Project finds a small but consistent excess of rare variants in binding sites for the CCCTC-binding factor which they conclude demonstrates a small deleterious effect for variation in binding sites for this transcription factor.[329] The ENCODE project found that 12% of trait-associated SNPs derived from the NHGRI GWAS catalogue overlapped a previously characterised transcription factor binding site versus 6% of total SNPs in 1000 genomes. This result suggests some binding site SNPs increase disease risk.[3] However, projects such as ENCODE have undertaken much of their work in cell lines. For our information to be truly disease relevant and to move from statistical association to specific molecular mechanism, some genomic assays need to be performed in the right cell type and at the right time. Furthermore, better *in vitro* and *in vivo* assays are needed to move from statistical associations to underlying mechanism.

We have run bioinformatic analysis to find disease-associated SNPs in binding sites for the Th1 and Th2 lineage master regulators T-bet and GATA3 in Th1

and Th2 cells. We have then established a new medium-throughput assay for testing these SNPs. We then moved to a mouse model system to try and better understand genomic regulation at a region highlighted by our analysis, the *IL18R1/IL18RAP* locus. We also investigated the role of one of the transcription factors, T-bet, in coeliac disease. Coeliac disease is one of the diseases associated with some of our hit-SNPs. In the course of our analysis, we set up a workflow for beginning to assay the functional relevance of some of the SNPs found by GWAS. Our workflow and some of the imminent additions to it are shown in 6.1.

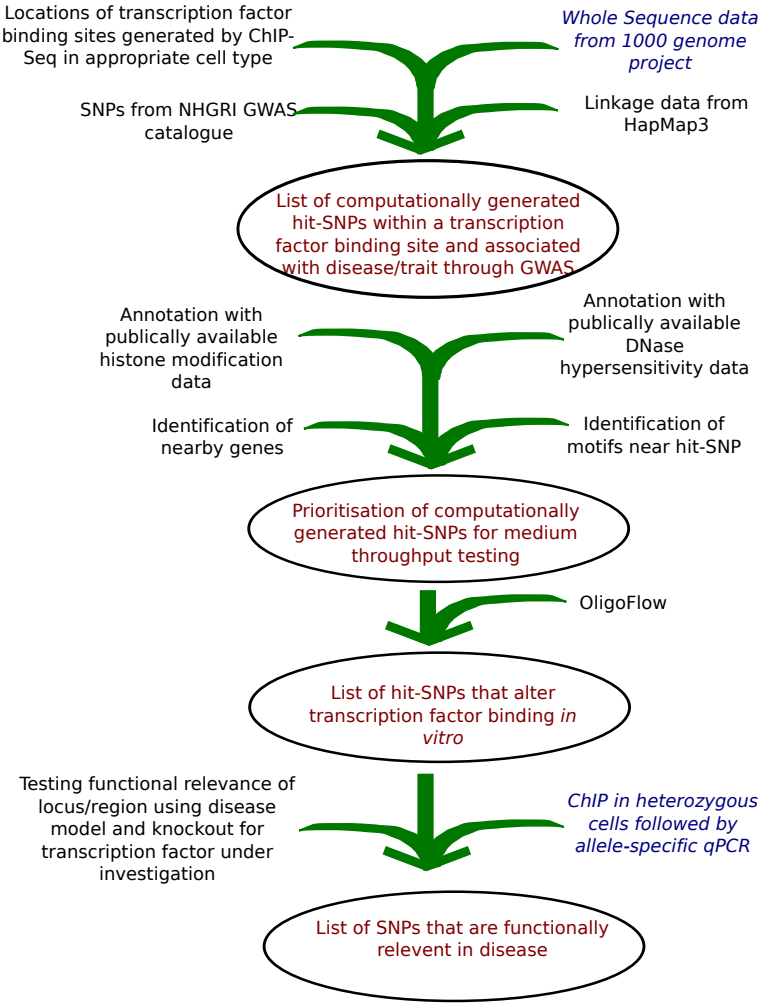

Figure 6.1: Workflow for project analysis. - Overview of the analysis performed in this thesis. Work performed is written in black. Analysis to be included in the near future is written in blue.

6.2 Genomic Annotation

6.2.1 Histone Modification and DNase Hypersensitivity

We found a total of 238 SNPs that were in a binding site for at least one of T-bet, GATA3 in Th1 cells or GATA3 in Th2 cells and were also in high LD with a SNP from the NHGRI GWAS catalogue. Many of these SNPs were found in transcription factor binding sites that intersected other genomic features such as the histone modifications H3K4me1 and H3K4me3 or regions of DNase hypersensitivity. An overview of features associated with our hit-SNPs can be seen in figure 6.2.

Our annotation was performed using publicly available datasets[10,303,304] and the intersections were performed on the Galaxy platform.[98–100] Therefore, very little computational expertise was required to annotate our SNPs and provide some prioritisation as to which SNPs we should test in our further assays. Further annotation will be possible when the newest datasets of the ENCODE project are no longer under embargo. Our annotation is similar to some of the annotation performed by projects such as ENCODE. However, we have the advantage that our ChIP-Seq data was generated in cells relevant both to the transcription factors under investigation and to disease pathology. One of the ENCODE project publications finds that a gene desert on chromosome 5p13.1 contains three SNPs rs9292777, rs11742570 and rs6451493 which are associated with Crohn's disease or ulcerative colitis and are in a binding site for GATA2 in HUVEC cells.[3] Of these SNPs, rs11742570 is also in a site of DNase hypersensitivity in both Th1 and Th2 cells. The authors hypothesise that this data might suggest functionally relevant GATA3 binding in Th1 and Th2 cells. In our study, we found that GATA3 does indeed bind at rs11742570 but only in Th1 cells, not in Th2. Closer examination of this region reveals that the peak of GATA3 binding at this SNP is quite small compared to a larger binding peak for T-bet and for GATA3 in both Th1 and Th2 which is approximately 20kbp upstream. Another SNP, rs929777, which is approximately 27kbp upstream of rs11742570 is in a binding site for T-bet but not GATA3 in either Th1 or Th2 cells. Both rs11742570 and rs929777 are found in a region of permissive histone modification. The summary data for these two SNPs can be seen in table 6.1. The SNP rs6451493 was not in a binding site for T-bet or GATA3. Although some of the generalities of these findings were predicted by ENCODE, the specific

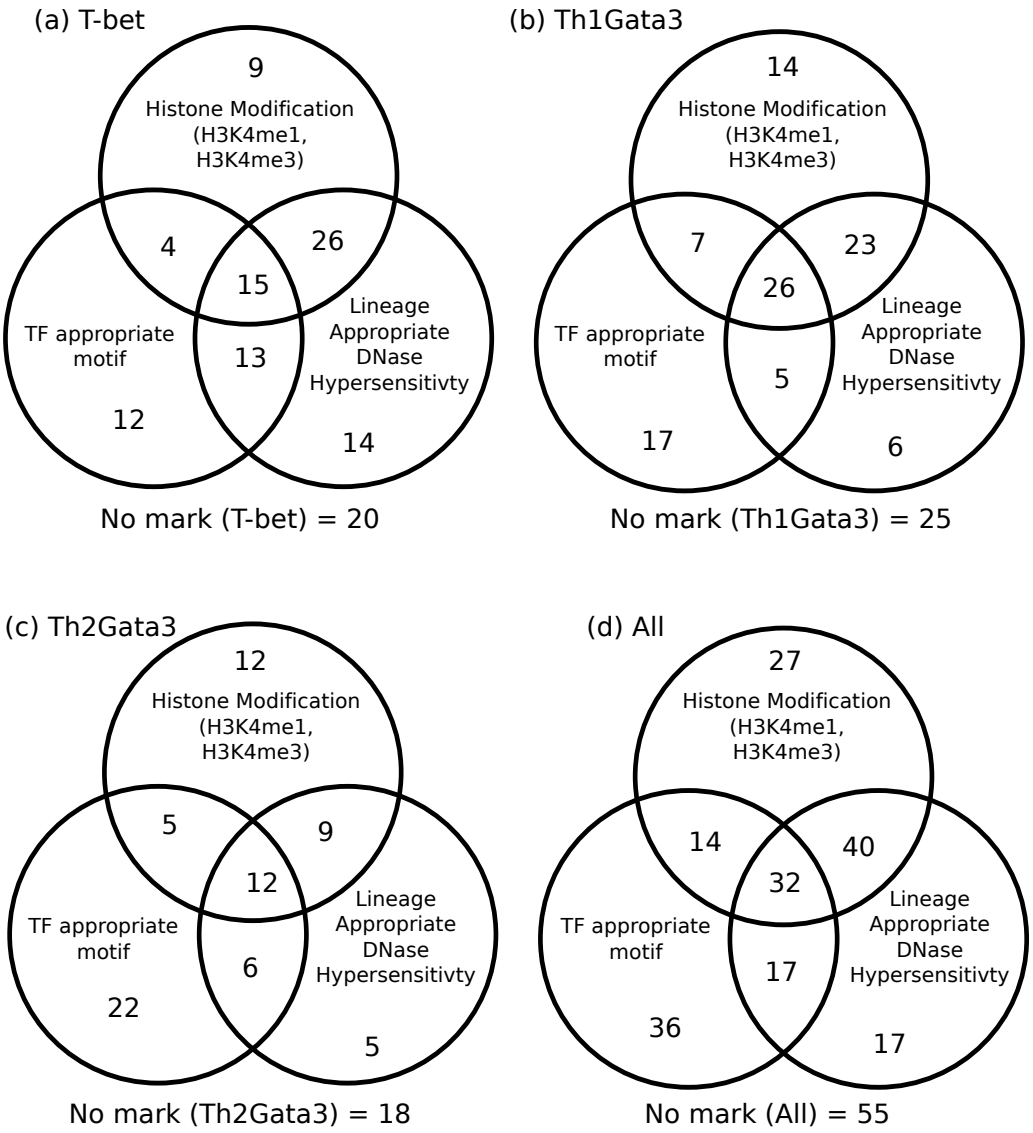

Figure 6.2: Summary of genomic features around hit-SNPs. - Histone modification motifs and DNase hypersensitivity sites found at hit-SNPs for T-bet (a), GATA3 in Th1 cells (b) and GATA3 in Th2 cells (c). Histone modification and DNase hypersensitivity also shown for all 238 hit-SNPs as a group (d). Only those SNPs in a region that matched the motif for the transcription factor under investigation (TF appropriate) are included in the motif count. Only those SNPs in a region of DNase hypersensitivity for Th1 cells (in the case of T-bet or GATA3 in Th1 cells) or Th2 cells (in the case of GATA3 in Th2 cells) are counted for DNase overlap (lineage appropriate DNase hypersensitivity).

binding patterns across different cell lineages was not.

Table 6.1: SNPs from ENCODE in T-bet and GATA3 binding sites. Genomic annotation of two SNPs that were hypothesised to be in GATA3 binding sites in Th1 and Th2 cells from ENCODE.

	Genomic Annotation rs11742570	Genomic Annotation rs9292777
Within Binding Site for GWAS SNP(s) in LD block	Th1Gata3 rs11742570, rs1373692, rs1992660, rs6451493, rs6896969, rs9292777	T-bet rs11742570, rs1373692, rs1992660, rs6451493, rs6896969, rs9292777
Associated Traits	Crohn's, ulcerative colitis, multiple sclerosis	Crohn's, ulcerative colitis, multiple sclerosis
Genes within 2kbp	none	none
Genes within 50kbp	none	none
Histone Modifications	H3K4me3	H4K4me1
DNase Hypersensitivity	yes	yes
Motif Present	no	no
Nonsynonymous SNP in high LD?	no	no

In our case, we wanted to use our genomic annotation to suggest which SNPs might alter transcription factor binding. We modified the traditional pulldown protocol to increase the accuracy of the results and decrease the time required for completion of the assay and make it medium-throughput. We ran comparisons between the traditional Western blotting pulldown protocol and our OligoFlow protocol and found good correlation between the results. However, due to the complex nature of genomic regulation and the novelty of both our observations and our technique we must be careful not to overstate our results at this time and work is ongoing on other ways to validate our findings. We are currently in the process of examining whether we can detect altered binding to our hit-SNPs *in vivo* using ChIP followed by allele-specific qPCR with TaqMan probes for the SNP of interest. However, we can draw some tentative conclusions from our work so far. The three hit-SNPs that did show convincing differential binding for T-bet by OligoFlow were rs1465321, rs1006353 and rs11135484. The details of these SNPs are summarised in table 6.2.

Table 6.2: Genomic annotation of hit-SNPs that showed differential binding in OligoFlow. Summary of genomic annotation for rs1006353, rs11135484 and rs1465321.

	rs1006353	rs11135484	rs1465321
Binding Site for?	T-bet, Th1Gata3, Th2Gata3	T-bet, Th1Gata3, Th2Gata3	T-bet, Th1Gata3
GWAS SNP(s) (r^2 with hit-SNP)	rs4771122 (0.81)	rs2549794 (0.83)	rs13015714 (1.00), rs2058660 (1.00), rs917997 (1.00)
Trait	Body Mass Index	Crohn's Disease	Crohn's Disease, Coeliac Disease
Genes with 0kbp	none	*ERAP2*(intron)	*IL18R1* (intron)
Genes within 2kbp	none	*ERAP2*	*IL18R1*
Genes within 50kbp	*MTIF3*, *GTF3A*	*ERAP2*, *LNPEP*	*IL18R1*, *IL18RAP*, *IL1RL1*
Modifications	none	H3K4me1 and H3K4me3	none
DNase Hypersensitivity	none	Yes	none
Motif Present (Algorithm used to find motif)	T-bet motif gained (`dreg`), GATA3 motif present (`dreg`), GATA3 motif altered (FIMO)	T-bet motif present (`dreg` and FIMO), GATA3 motif present (`dreg`)	T-bet motif present (`dreg`)
Other Motifs Present	Evi1 (TRAP)	Partial Runx3 (MEME), Runx1, SpI1, IRF1, FEV, ELV5, IRF2, Egr1 (TRAP)	AP1, FEV, ELV5, NFE2L2 and RORA_2 (TRAP)
Binding site also called by SiSSRs	T-bet site called. GATA3 sites (Th1 and Th2) not called	T-bet site called. GATA3 site called in Th2 cells only	T-bet site called. GATA3 site not called.

All three SNPs were in binding sites for both T-bet and GATA3 in Th1 cells. Although rs11135484 was associated with histone modifications and DNase hypersensitivity, these were not universal features of all our hit-SNPs and were not found around either rs1006353 or rs1465321. Given that neither rs1006353 nor rs1465321 are close to a TSS, the lack of histone modification is not surprising from a genomic perspective. H3K4me3 is known to mark promoter regions. Although H3K4me1 is often treated as a mark of enhancer regions the paper from which our data was taken shows that H3K4me1 still tends to be found within 2kbp of a TSS.[10] The anomaly, by this reasoning, is rs11135484 which is further than 2kbp from the TSS for *ERAP2* but yet is associated with both H3K4me3 and H3K4me1 modifications. This result suggests that the presence or absence of a modification cannot predict altered transcription factor binding, although whether it can distinguish those differential binding events that affect phenotype from those that do not remains to be investigated. It should be noted that the Barski *et al*[10] paper produces data on resting T cells and not cells that have been activated and then skewed towards a specific lineage. As such the cell type used to annotate our data with histone marks is not perfectly matched with the cell type which we used for our binding site data. However, the data from Barski *et al* is the closest data set available. For this reason, we decided not to repeat the analysis with published acetylation data. As mentioned in the introduction, methylation is a fairly stable mark that indicates areas that are generally transcriptionally permissive whereas acetylation is a less stable mark that indicates active transcription of the gene. Given the differences in active transcription between resting $CD4^+$ cells and skewed Th1 and Th2 cells we decided that published data did not match our data sufficiently for any analysis of acetylation marks to be useful. Data on the histone modifications should be interpreted with this limitation in mind.

The lack of DNase hypersensitivity at rs1006353 and rs1465321 is slightly unexpected as transcription factor binding tends to loosen DNA and result in DNase hypersensitive sites. Furthermore, the DNase hypersensitivity data was as matched as possible to our own in terms of cells studied. Given the biology of Th1 cells, it is surprising that a SNP in an intron of the *IL18R1* locus does not overlap a region of DNase hypersensitivity. However, if we examine the genomic location of rs1465321 (fig. 4.12) we find that it is just upstream of a region of differential DNase hypersensitivity between Th1 and Th2 cells. The differences in DNase hypersensitivity at this region between the two cell lin-

eages most likely reflects the different genomic architecture of the locus in Th1 versus Th2 cells. Thus rs1465321 is near an important region but our first pass analysis for DNase hypersensitivity was too blunt a tool to capture this. This result highlights the need to narrow down and focus on specific genomic areas guided by other data and knowledge. Simply overlapping transcription factor binding sites with other features could miss important sites.

A further finding in the *IL18R1* locus is that rs2058622 is in a DNase hypersensitive site in Th2 but not Th1 cells. This SNP is in a binding site for GATA3 in Th1 and Th2 cells and lies between two T-bet peaks. The difference in DNase hypersensitivity patterns at the *IL18R1* locus demonstrates the importance of using the most appropriate cell type possible for genomic annotation. In our analysis, hit-SNPs for bipolar disorder (rs7578035), antipsychotic response (rs507101), height (rs10152591), ulcerative colitis (rs1886730), schizophrenia (rs2275271), cardiographic conduction (rs4679048), conduct disorder (rs17103930) and quantitative traits (rs3758253) are all found in a binding site for T-bet or GATA3 in Th1 cells and a binding site for GATA3 in Th2 cells. However, they are only in a DNase hypersensitive site in Th1 cells. Hit-SNPs for asthma (rs12991737), white blood cell types (rs3732123) systemic sclerosis, leprosy and chronic lymphocytic leukaemia (rs9271612 and rs9271613), bone mineral density (rs851984), coeliac and Crohn's diseases (rs2058622) and basal cell carcinoma (rs4772190) are all found in a binding site for T-bet or GATA3 in Th1 cells and a GATA3 binding site in Th2 cells but only in a DNase hypersensitive site in Th2 cells. It is unclear why the DNA around some of these SNPs may be selectively loose in a lineage specific way. Some of these differences may result from experimental noise generated at one of the many data acquisition or analysis steps required to produce the information. For example, it is interesting to find Th1 specific DNase hypersensitive sites around SNPs associated with various psychological conditions. However, while this possibly merits further investigation, it is also likely that these SNPs are false positives generated by the large number of GWAS conducted for psychological conditions. The ulcerative colitis associated SNP, rs1886730, which is in a binding site for GATA3 in Th1 and Th2 cells but only in a DNase hypersensitive site in Th1 has greater support from evidence outside of genomics. The interplay between Th1 and Th2 cells in the inflammatory bowel diseases is well known. This SNP is in an intron for *TNFRSF14* (also know as HVEM or LIGHTR) which has been shown to be protective in an adoptive transfer model of colitis.[330] H3K4me3 and H3K4me1 histone modifications

are both present and the SNP alters a GATA motif. The Th2 specific DNase hypersensitivity at rs9271612 and rs9271613 which are associated with systemic sclerosis, leprosy and chronic lymphocytic leukaemia also correlates with non genomic data. Systemic sclerosis is an autoimmune condition and both Th1 and Th2 cells have been implicated in protection from and susceptibility to various forms of cancer. The link to leprosy is also interesting. Although leprosy is caused by infection with *Mycobacterium leprae*, the pathology varies between individuals and this variation correlates with a type I or type 2 response from $CD4^+$ and $CD8^+$ cells.[331] These SNPs are in a binding site for GATA3 in Th1 and Th2 cells and within 50kbp of the HLA locus suggesting an immediate link with the immune system.

6.2.2 Genes of Interest- Immune Genes

We found many genes known to be of importance in the immune system within 50kbp of our hit-SNPs. SNPs in the region of *ERAP2* and the *IL18R1/IL18RAP* locus have already been discussed. However, we also found hit- SNPs in or within 2kbp of *TAGAP* (rs1738074, T-bet), *IL10* (rs3024505, T-bet), *RUNX3* (rs4265380, T-bet), *CXCR2* (rs6723449, T-bet), *RUNX1*(rs8129743, T-bet), *TNFRSF14* (rs1886730, Th1Gata3, Th2Gata3), *ICOS* (rs4522587, Th1Gata3, Th2Gata3) *GZMB* (rs8192917, Th1Gata3) and *ICAM4* (rs281438, Th1Gata3 and Th2Gata3). When we searched for genes within 50kbp of our hit-SNPs, we also found *ATG16L1* (rs10929322, T-bet, Th1Gata3, Th2Gata3), *CTLA4* (rs11571293 for T-bet or rs231727 for Th1Gata3, Th2Gata3), *IL13* and *IL4* (rs12653750, T-bet), *IL12A* (rs485499 and rs485789, T-bet), *IL2RB* (rs743776 and rs743777, T-bet), *IL7R* (rs4024110, T-bet, Th2Gata3) and *NOD2* (rs8062727, T-bet, Th2Gata3). We also found many SNPs in binding sites for T-bet and GATA3 that were within 50kbp of the HLA locus.

For the T-bet binding site hits, rs743776 and rs8062727 have already been tested and show no altered binding by OligoFlow. From our first round of OligoFlow results we hypothesise that most of the remaining SNPs will show altered or even any binding as there is no T-bet consensus sequence present. The exceptions are rs485789 and rs743777 which were shown to have a T-bet motif by the FIMO algorithm. Testing whether these SNPs show altered binding will give us an indication of the relative capacities of `dreg` and FIMO to predict SNPs that

might alter binding. So far, hits in the OligoFlow analysis have arisen from regions containing a T-bet motif present as called by the `dreg` algorithm. Of the regions around rs1006353, rs11135484 and rs1465321, only the region around rs11135484 contained a strong T-bet motif by FIMO analysis.

In more general terms, our results provide a convenient way of narrowing down large scale genomic data to generate specific testable hypotheses. GWAS hits highlight regions where small genetic variations can influence the immune system and disease risk. We could hypothesise that the expression of genes in these regions would be more easily perturbed than near loci with no GWAS SNPs. Furthermore, we might suggest that small perturbations in the expression of these genes would have a greater influence on disease risk than small perturbations in expression of genes with no GWAS SNPs. We would expect a causal SNP to alter either transcription product or genomic architecture. In the case of genomic architecture, we might expect both T-bet and GATA3 to have a role in shaping any locus to which they bind. Therefore, we could hypothesise that any hit-SNP implicates a role for T-bet and/or GATA3 acting through that locus in the GWAS trait, whether the SNP directly alters binding or not. It is interesting to note, for example, that T-bet binds over a SNP within 50kbp of IL-4 and IL-13.

6.2.3 Genes of Interest - Non Immune Genes

To focus exclusively on immune system genes would be to neglect other potentially important SNPs. A case in point is the SNP rs1006353 which is not near any immune genes. However, rs1006353 is within 50kbp of Mitochondrial Translation Initiation Factor 3 (MTIF3) and is associated with Body Mass Index. The role of mitochondria in energy harvest is well documented and a role for T-bet in obesity has been published.[332] Perhaps T-bet has a role in regulating MTIF3 expression in some cell types. As we see with the IFN-γ locus, it is possible for genes to be regulated by different mechanisms under different circumstances.[170]

Our analysis returned a few coding regions about which very little is known as highlighted by the use of an 'orf' or 'LOC' classification. Given the background to and rational behind our work we might hypothesise that some of these regions could be involved in the immune system. In support of this idea,

a role for C12orf52 in Notch signalling has recently been published.[333] C12orf52 is within 2kbp of rs28365932 a SNP in a T-bet binding site. As discussed in the introduction, Notch signalling has a role in T helper cell lineage commitment. We could investigate a potential role for this and other regions in $CD4^+$ cells in a fairly straight forward manner by transduction and cytokine measurement in cells skewed Th1 or Th2 as appropriate.

Our analysis also returned SNPs near non protein coding genes including miRNAs and a cluster of small nucleolar RNAs within 50kbp of rs507101 (Th1Gata3, Th2Gata3). Of particular interest is miR-4772 which is located within an intron on *IL18RAP*. This miRNA is within 50kbp of a few of our hit-SNPs including rs1420106, which is in the promoter region of *IL18RAP*. Although mir-4772 is not within 50kbp of rs1465321 it is part of the *IL18R1/IL18RAP* locus. Given the dynamic nature of this region and the role of T-bet in genomic structure of this region, we might hypothesise that this miRNA also has a role linked to immunity. To date, very little is published on the role of this miRNA.

We also found SNPs that were not within 50kbp of any genes. However, the presence of other features at these SNPs suggest the regions are important in genomic regulation. For T-bet 18.6% (21 in 113) hit-SNPs were not within 50kbp of a gene. For GATA3 in Th1 cells 13.0% (16 in 123) hit-SNPs were not within 50kbp of a gene and for GATA3 in Th2 cells 10.1% (9 in 89) SNPs were not within 50kbp of a gene. Yet, as seen with the SNPs rs11742570 and rs9292777 in the gene desert on 5p13.1, there are other genomic features present at these SNPs. Figure 6.3 shows the percentage of each set of hit-SNPs that are not within 50kbp of a gene but have marks. Multiple marks are seen at some SNPs. The presence of chromatin marks suggests that these SNPs do have a role in gene regulation. By comparing their associated trait with the known roles of our transcription factors we could start to form hypotheses as to the regulatory role served by the regions containing these SNPs.

6.2.4 Co-binding

The importance of co-binding is highlighted by the overlap of SNPs in T-bet and GATA3 binding sites in Th1 cells. As discussed in the introduction, co-binding of transcription factors, including co-binding of the master regulators,

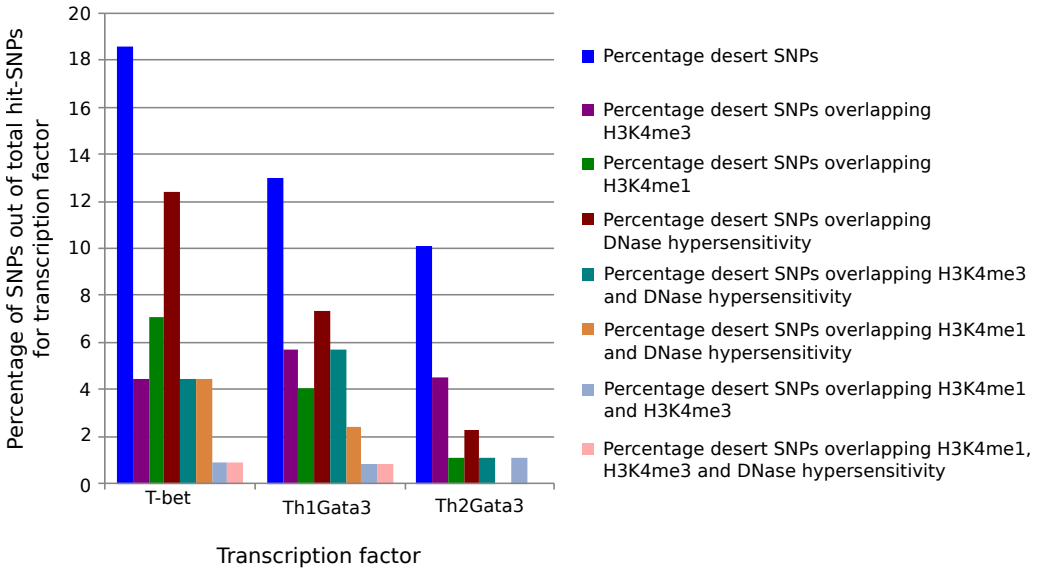

Figure 6.3: Annotation of SNPs not within 50kbp of a gene. - Percentage of hit-SNPs that are not within 50kbp of a gene (desert SNPs) and percentage of desert SNPs at locations of histone modification and DNase hypersensitivity.

is an important aspect of control within the immune system. We found many SNPs that were in a binding site for both T-bet and GATA3 in Th1 cells which is in agreement with recently published work.[234] Our top three hit-SNPs from the OligoFlow, rs1465321, rs1006353 and rs11135484 were all in a binding site for T-bet and for GATA3 in Th1 cells. The SNPs rs1006353 and rs11135484 were also in a binding site for GATA3 in Th2 cells. By contrast, rs1465321 was not in a binding site for GATA3 in Th2 cells, possibly reflecting the changes at the *IL18R1/IL18RAP* locus between Th1 and Th2 cells. Of interest, there are a few SNPs that are in a T-bet binding site in Th1 cells and a GATA3 binding site in Th2 cells but not in a GATA3 binding site in Th1 cells. The binding of T-bet in Th1 and GATA3 in Th2 might suggest that the SNP is in a region that is differentially regulated in Th1 versus Th2 cells. The five SNPs in this category are shown in table 6.3.

Table 6.3: Hit-SNPs for T-bet and Th2Gata3 but not Th1Gata3. Details on SNPs that are in a binding site for T-bet and for GATA3 in Th2 cells but not in a binding site for GATA3 in Th1 cells.

Hit-SNP	Associated-Trait	Genes within 50kbp	Histone	DNase	Motif
rs4024110	Ulcerative Colitis	CAPSL, UGT3A1, IL7R	none	Yes	none
rs727263	Basal Cell Carcinoma	UBAC2, mir-623, GPR183	none	Yes	T-bet(P.D), GATA3(P.D)
rs6556405	Mean Platelet Volume	RNF145	Promoter	Yes	none
rs8062727	Leprosy	NOD2, SNX20, NKD1	Enhancer	Yes	T-bet(A.D)
rs2984920	Coeliac Disease	RGS1	Promoter	Yes	T-bet(P.D)

All of these SNPs are in DNase hypersensitive sites and most have motifs and/or are for traits with an immune component. Of note, rs8062727 has already been tested in OligoFlow and showed no differential binding in the first experiment. Although the SNP is in a fairly strong consensus motif, we did not see differential or even any binding above the negative control at this SNP. This may be for one of two reasons. T-bet may not bind at that location either because binding is to another region within the ChIP-Seq peak or because the peak is a false positive in the ChIP-Seq data. Alternatively, the OligoFlow result could be a false negative because it does not provide the appropriate co-binding transcription factors or the appropriate chromosomal context. In support of the first idea, rs8062727 was not in a T-bet peak as called by the SISSRs algorithm whereas rs1465321, rs11135484 and rs1006353 were in peaks called by both MACS and SISSRs. Furthermore, our experiments suggested that YT cell lysate would produce the same results as lysate from Th1 cells. In support of the second idea, rs8062727 is in the Brachyury half site of the T-bet motif which does suggest that it should alter binding.

The SNP rs2984920 has also been tested in both OligoFlow and pulldown followed by Western blotting. In both techniques the results were mixed. One of the Western blots and one of the OligoFlow experiments showed increased binding at the G allele of the SNP. However, the other OligoFlow experiment

showed little differential binding and a second Western blot experiment showed more binding at the A allele (figs. 4.1 and 4.9). It is possible that the differential binding of GATA3 at this locus between Th1 and Th2 cells highlights a greater requirement for the correct cofactors to bind at this locus. This SNP and the other four SNPs listed in table 6.3 would provide a good starting point for a thorough examination of differences in transcription factor binding in the OligoFlow assay when using YT cell lysate versus Th1 cell lysate.

The SNPs listed in table 6.3 are mixed with respect to the presence of a T-bet binding motif. In some contexts, T-bet may bind DNA through other transcription factors so the T-bet motif is not seen but other lineage specific transcription factors are required. This raises the possibility that OligoFlow with YT cell lysate might only demonstrate altered binding at SNPs near or within a motif, not because a motif is always present but because situations where a motif is not present require other cofactors that are present in Th1 cells and not in YT cells.

To ultimately resolve these issues, other techniques are needed. In this project, we established a luciferase construct for testing hit-SNPs and used it to analyse altered gene activation by the two alleles of rs1465321. However, this assay has several issues. From the small MFI differences seen in the OligoFlow assay, we expect any difference in luciferase gene activation between the two alleles of a SNP to be subtle. Given the potential for variability in electroporation efficiency we could raise concerns over whether the assay is sensitive enough to detect such subtle variations reliably. Although we used a renilla luciferase transfection control plasmid, our experiment was based on the assumption that this plasmid entered the sample with equal efficiency to the promoter construct and the expression plasmid. Although this assumption is widely made and broadly accurate in most luciferase assays, it may not be accurate enough for our purposes. Accuracy could be improved by using a plasmid containing both renilla and firefly luciferase genes on the same vector. However, this might then raise issues over whether two genes on the same plasmid are regulated independently. Furthermore, this method does not remove the assumption that the efficiency of transfection of expression plasmid is proportional to that of the luciferase vector. Furthermore the assay, as with the OligoFlow, may not provide the correct cofactors. The experiments were performed in EL-4 cells which are derived from a murine lymphoma. EL-4 were used as they do not express

T-bet and therefore provide a clean background into which T-bet can be transfected in a controlled way. EL-4 cells are also transfected easily to high levels. However our constructs were based on the human genome and, as previously discussed, gene activation may require other transcription factors in addition to T-bet. Some of these factors may not be sufficiently provided by the murine EL-4 cells. We did try the experiments in CEM cells, in addition to EL-4. CEM cells are a human T cell lymphoblast cell line. However, the transfection efficiency into these cells, while good, was not good enough to detect the subtle differences we were expecting.

Furthermore, the luciferase assays provide a larger section of DNA around the SNP than the OligoFlow assay but still do not test the SNP in a full native chromatin environment. Luciferase assays have been found, by others, to be less able to demonstrate transcription factor binding at regions distant from the TSS.[27] Because many of our SNPs were not near a TSS, we decided to make a construct containing a promoter known to be activated by T-bet, the *IFNG* promoter. We hoped we could then test for a differential ability of our SNPs to act as enhancers or otherwise on this promoter. However, this still does not provide a native context in which to test altered binding and its functional effects.

6.2.5 Motifs

A transcription factor binding motif was found in the region around all three of our strongest hit-SNPs, rs1006353, rs11135484 and rs1465321, using the `dreg` tool (fig. 6.4.) However, this motif is only altered for rs1006353. For both rs1465321 and rs11135484, the motif is close to but not altered by the hit-SNP. By contrast, other hit-SNPs that do alter a consensus motif did not show altered binding by OligoFlow. The sequences around those SNPs that were found to alter a motif and a summary of the results from the OligoFlow work are shown in table 6.4.

rs1006353

CAGAGGCCTTTGGGGGTTGGCTGGAGCAGCACCACCCCTC
TGTGGTGTGCACCTGGAGCACCCATCCACATAGGCCCCA
[A/*G*]
TGATAAGATAGCGTTAACCACCAGCTGTGCTGTAGCAACC
CAGCCCCTGTGGATGGCCACATAGTGGGTGAAGAAGTCCC

rs11135484

CATGTGATCTCCAACAATTCTGTGAACATTTCAGAGTCTC
TGTTTCCTCACCTGAGAAACAACACCAACC***TCACAC***CCAC
[G/A]
TAACAGGATTAAAAGATAATGTGCAGCCTCTAGTTCAGTT
TCACTTCCTGTTTTCTTTTTCCACAGGGGTGTACTTCTTG

rs1465321

CTCCATGTGCAGGACACTGCTCTGGGAACACTGAAACAAT
CAACAACACACTGCACCTGCCATCCAGGAGCTCCAAGGCT
[G/A]
GGG***TTAACAC***TGAGGCCAACTGACCCAAAGCTGTGACCAC
ACTTTCCTGCTGCAGGCCCAGCCCATCCTGGTTATGAAAA

Figure 6.4: Surrounding sequence for hit-SNPs. - Sequence around rs1006353, rs11135484 and rs1465321 is shown. Red text indicates alternative alleles at SNP. Text in blue italics indicates consensus motif for T-bet. Text underlined in green is GATA consensus motif.

Table 6.4: Summary of OligoFlow for SNPs which altered T-bet motif. Sequence around hit-SNPs tested in OligoFlow where hit-SNP altered T-bet consensus motif and summary of results from OligoFlow. Opposite alleles of each hit-SNP are given in bold. Match to consensus motif is given underlined.

Hit-SNP	Sequence around SNP	Summary OligoFlow Experiments
rs10152590	GTGAGGAGGTGGGAGGAA-CATTCATG**[A/T]**GTGGGTAA-CCCTGTGACTAATCCAG	Differential binding seen but in opposite directions over two experiments. Not consistent.
rs2387397	ATGCAAATAAGAAGCTGTT-TCAGTGT**[C/G]**TGCCCATCT-GAGACGCTGACATAAA	No differential binding seen in first experiment.
rs2703078	GCTATGACCACTGGCTCAC-ATTTGCC**[A/G]**CTTCAAGCC-TCCTACAAGAAACATA	Consistent differential binding seen across 2 experiments but very small in magnitude.
rs5778	ACTCTTCCAGCCTCCCACA-TGATGGG**[C/T]**GGAAAAAG-GCAAAAGCCCAGATTAA	No differential binding seen in first experiment.
rs743776	ATCCAAACTCTCAGTCTTG-ACCCACA**[C/T]**GTCCTACAG-GGACTGCCCCATCTTC	No differential binding seen in first experiment.
rs7441808	TATTTACCTTAGGGACTCCT-CTGGGT**[A/G]**TGTGAAGAA-TTCCCCTGTTTTGCTC	No differential binding seen in first experiment.
rs8062727	CCACAGGGGAAAAATGTG-TGGTTGCC**[A/G]**CCACTTCC-TCTTATGGGGAAAGGAG	No differential binding seen in first experiment.

We could argue that the GTGTG/CACAC motif found around rs10152590, rs2387397, rs743776 and rs7441808 was not a good representation of the T-bet motif because it lacks a further A/T at the end of the motif which is highly conserved in the PWM.[234] This could explain why we do not see differential binding around these SNPs. CACAC is found near rs11135484 but the presence of the further, highly conserved T (fig. 6.4) may distinguish this site. The SNPs rs2703078, rs5778 and rs8062727 are better matches to the consensus motif with both the Brachyury half site and the conserved A/T. In each case, the SNP alters the central nucleotide of the Brachyury half-site (GTG or CAC) which is

highly conserved in the motif, yet the impact on binding is small or non existent. By contrast, rs1006353 alters the first nucleotide of the Brachyury half site which is less conserved and yet rs1006353 does alter T-bet binding in the OligoFlow assay. This finding highlights the issues with drawing conclusions about SNP effects from motifs. The situation is not improved by using a PWM in the bioinformatic analysis: FIMO found a T-bet PWM match in the region around rs11135484 but not the regions around rs1006353 or rs1465321.

It is possible that some instances where no differential binding is found could result from using YT cell lysate and not lysate from Th1 cells. It is possible that in some cases a necessary binding partner is absent in the YT lysate. However, for our strongest hit-SNPs the trend in results did not differ between using lysate from YT and lysate from Th1. Our results demonstrate that an altered motif does not correlate with altered binding in OligoFlow. However, so far our results do suggest that binding is not seen in DNA sequences that do not match to any consensus motif. However the region around rs8008961, which does not contain a consensus sequence has, so far, given conflicting results and needs to be retested. The SNP rs8008961 is in an intron for *RAD51B* and is associated with Primary Biliary Cirrhosis.

6.2.6 Nonsynonymous SNPs

Only 20.6% (49/238) of our hit-SNPs were generated by a trait-associated SNP that was also in high LD with a nonsynonymous coding SNP. This result demonstrates, in agreement with many large scale genomic projects, that many of the trait-associated SNPs in which we were interested, were not acting through altering protein structure. We prioritised testing those SNPs that were not part of an LD block containing a nonsynonymous SNP for setting up our *in vitro* testing. However, we cannot assume that these hit-SNPs do not exert a mechanistic effect through altering transcription factor binding and, as we establish further assays, some of these SNPs should be tested.

6.2.7 Future Directions for Genomic Annotation

An obvious extension to our *in silico* analysis is the inclusion of recently published data from the 1000 genomes project.[329] For our project we had already

established the pipeline for our *in silico* analysis and moved on to establishing *in vitro* and *in vivo* testing before the data from the 1000 Genome Project became readily available. However, the increasing use and availability of whole sequence data make this an important next step for the type of analysis performed here. The authors of the 1000 Genome publication analyse the number of SNPs in LD (r^2>0.5) with each GWAS signal in the NHGRI GWAS catalogue and find more than two times as many SNPs using 1000 genome compared to HapMap. This increase could greatly help us to find other transcription factor binding SNPs. Many of the novel SNPs found in 1000 genomes are rarer SNPs (Minor Allele Frequency < 5%) which raises the question of how likely the new SNPs are to be causal in common diseases. The debate over the relative contributions of rare versus common variation in the common complex diseases is still active in the genomics community.

In our current analysis, we used the HapMap3 CEU population for all LD calculations even though not all GWAS are performed in a population of CEU-like ethnicity. Rerunning our analysis with other populations might yield more data or more appropriate data. For example, we have discussed the disease leprosy: GWAS on leprosy have often been performed in populations of Asian ethnicity. If we expand this analysis to include the rarer SNPs of the 1000 genome data, this issue becomes more important as rarer variations have a greater tendency to only be found in some ethnicities.[329]

Moving on from bioinformatic analysis, it is important to be able to test for altered transcription factor binding and downstream effect both *in vitro* and *in vivo*. To this end, we developed the OligoFlow assay. The OligoFlow assay allows altered binding between two different alleles to be tested quickly and should, in theory, be easily extendable to other transcription factors. In practice, it is likely to work better for some transcription factors than others as demonstrated by our work on GATA3 binding. In the case of GATA3, the strength of binding was heavily dependant on the sequence surrounding the positive and negative control motifs making designing positive and negative controls difficult. By contrast, for T-bet our positive and negative controls worked well. This difference may reflect the difference in promiscuity between the two transcription factors and imply that this assay is more suitable for those transcription factors with a longer and better defined consensus binding sequence. Examining the potential use of the OligoFlow assay for other transcription factors

would be a useful piece of further work. For ease of analysis such work should, from the outset, use a fixed number of cells and fixed cytometer settings. This would make replicate experiments, acquired over time, more easily comparable and allow for better statistical analysis of samples. In our case, this was not possible as we were optimising the assay at the same time as collecting data and had limited time for repeating measurements.

Another way of testing for altered binding, in addition to the OligoFlow assay would be desirable. In this project, we used a luciferase assay to investigate potential differences in binding at rs1465321. However, this assay was not very sensitive and was subject to large experimental variation due to the use of electroporation. Work to test for differential binding in the appropriate genomic context is ongoing. We plan to test altered binding by cross-linking Th1 cells from individuals known to be heterozygous at the SNP of interest. We then plan to perform ChIP on these cells followed by allele-specific qPCR to examine for enrichment of one allele over the other in the fragments that are pulled down. We hope this assay will be sensitive enough and will account for the various trans and cis acting factors which affect transcription factor binding, but which may be absent in the OligoFlow assay. Such a technique would allow independent validation of the OligoFlow assay. Ultimately, we would like to establish a pipeline in which *in silico* hit-SNPs are screened by the medium to high throughput assay of OligoFlow and then SNPs that show differential binding in OligoFlow are further tested in heterozygotes by ChIP and allele specific qPCR. The ChIP step is likely to be time consuming as sufficient heterozygotes for the SNP under investigation must be found. Then cells from each heterozygote would have to be grown and subject to ChIP. Thus, the screening out of false negatives from the bioinformatic analysis by the OligoFlow assay would be highly valuable in reducing the number of false negatives from the *in silico* analysis prior to further testing.

Another important avenue to pursue further would be to analyse some of our SNPs for eQTLs in the appropriate cells types. In our case, we should examine some of our SNPs and their effect on gene expression in skewed and stimulated Th1 and Th2 cells. The importance of this was demonstrated in a recent IBD meta-analysis which found that SNPs associated with Crohn's diseases and ulcerative colitis are found near genes that show cell specific expression most notably in DCs but also in $CD4^+$ cells.[279] Analysing gene expression with re-

spect to genotype in the distinct cell populations of $CD4^+$ cells skewed to Th1 or Th2 would be fairly straight forward on the relatively small number of SNPs and genes identified by our *in silico* analysis.

6.3 Immune Mediation

T-bet directs aspects of the immune system and its dysregulation results in immune related conditions. We wanted to investigate whether our hit-SNPs were enriched for SNPs in LD with GWAS SNPs for immune related conditions. The GO ontology provides a useful means of describing genes and gene products and classifying them based on their characteristics. No such ontology exists for phenotypic traits or diseases and for many traits and diseases any classification system would have controversial elements. As such, it is difficult to objectively examine whether a list of traits is enriched for immune mediated conditions. We attempted to solve this problem by constructing lists of definitely, possibly and definitely not immune related conditions based on majority clinical judgment. Although some conditions on each list might be open to debate, this provided a useful starting point to try and analyse our SNPs for immune enrichment in the same way that a list of genes might be analysed for enrichment of genes with a certain GO characteristic.

Our analysis showed significant enrichment for strictly immune related hit-SNPs in T-bet binding sites and significant enrichment for both strictly and loosely immune related hit-SNPs in GATA3 binding sites in Th1 cells ($p < 0.05$ in all cases.) If we take the average number of strictly immune related SNPs found in 10,000 randomly shifted T-bet sites versus the actual number of strictly immune related hit-SNPs then this suggests that over 25% (11/42) of our strictly immune related T-bet hits are functionally relevant over the number of SNPs expected by juxtaposition of immune related SNPs and immune related genes. Although we would need to test all hits by other methods to determine how accurate this prediction was, the result is encouraging. However, we must be aware that such a classification has the potential to follow a circular argument. We were interested in whether hit-SNPs for T-bet were enriched in immune related traits. However, we could argue that the involvement of T-bet implicates the immune system in a trait. Researchers working on T-bet might be more likely to judge a borderline condition immune because T-bet is known to have

role. To avoid this, labels were assigned to each condition by clinicians independently from each other and from our project. However, given the research environment in which the question was asked, we cannot guarantee that such bias was completely removed. A future direction of this work, might be to seek consensus from a larger group. It is important to note that such enrichments for immune related SNPs were weaker for hit-SNPs in GATA3 binding sites in Th2 cells often failing to reach significance. This may reflect the fact that GATA3 has roles beyond the immune system. Although our ChIP-Seq data was from Th2 cells, it is unlikely that all the binding sites were T cell specific. More general GATA3 binding sites with a role in other systems could be more common in Th2 cells than Th1 cells where GATA3 expression is lower and GATA3 is redistributed across the genome by T-bet.[234] Alternatively, this could reflect ambiguity in classifying diseases as immune mediated. Aberrant Th1 responses may produce disease with a more obvious immune component than aberrant Th2 responses.

We could envisage a situation where transcription factors known to have a prominent role in the immune system might assist in establishing a list of traits that could then be used in further analysis. To some extent, the merits of the classification that we made were verified by the permutation and randomisation testing that we performed. We used data on binding sites for ER-α, a transcription factor with little or no role in the immune system and NF-κB, a transcription factor known to have a key role in the immune system. We showed that random reassignments of our classification produced significantly fewer immune mediated hits than our actual classification. This was done independently of work on T-bet and GATA3 binding sites. Choosing transcription factors that are or are not involved in the immune response can be done through the GO ontology. This could highlight an immune component to conditions not previously considered immune mediated. A case in point could be the juxtaposition of an established immune component of obesity, the recent publication of a role for T-bet in obesity[332] and the finding that one of our strong hit-SNPs is associated with BMI.

The enrichment of immune mediated conditions in our hit-SNPs for T-bet provided a cheap and relatively quick way of gaining some confirmation as to the validity of our *in silico* analysis. However, there are issues to consider. Many complex traits are heterogeneous with respect to the extent and mechanism of

immune involvement. In some cases this heterogeneity is recognised and might be discussed in the GWAS publication. In other cases such heterogeneity is not fully understood or appreciated. Either way, our analysis was too broad to capture finer details of each published study. Furthermore, even dividing transcription factors into immune and non immune related is not a clear cut process. For example, the data that we used on ER-α was taken from a paper which demonstrated that, in some cases, ER-α can be tethered to the genome by Runx1. Our binding site data for ER-α, therefore, would have included cases of Runx1 binding sites where ER-α was tethered and not just direct ER-α DNA binding events. Although ER-α is not immune related, Runx1 does have a role in the immune system. As discussed in the introduction, Runx1 is required for activation of the Th17 master regulator RORγT and can be sequestered from this role by physical interaction with T-bet.

6.4 The *IL18R1/IL18RAP* Locus and rs1465321

6.4.1 Possible Mechanisms

To investigate a potential role for altered binding at one of our hit-SNPs, rs1465321, we examined the kinetics and T-bet dependence of IL18R1 expression *in vitro* and *in vivo*. Fine mapping data, using the densely covered ImmunoChIP has been published for coeliac disease and Crohn's disease, the two diseases associated with rs1465321.[75,279] It is important to note that both of these studies find other candidate SNPs with greater statistical significance than rs1465321. For coeliac disease, the highest association in the region was found at rs990171, a SNP downstream of *IL18RAP*. For Crohn's disease, the greatest significance was found for rs6708413, which is in an intron of *IL18RAP*. These results, coupled with eQTL data concerning IL18RAP implicate this locus far more prominently than the *IL18R1* locus. However, it is possible that an eQTL for IL18R1 might only be observed in the functionally relevant Th1 cells rather than whole blood or bulk PBMC. Because of the limitations of our OligoFlow and luciferase assays and because work on other assays is still ongoing, we cannot conclude any definite role for rs1465321. However, our annotation of the *IL18R1/IL18RAP* locus combined with the importance of this signalling pathway in immunology suggested that a further study of this locus, in the context

of T-bet and GATA3 binding and disease mechanism, would be useful. This is especially the case given that the ImmunoChIP data for IBD demonstrates that genes near the hits discovered for IBD to date are heavily enriched for genes involved in cytokine, and in particular IFN-γ, production.[279]

In terms of kinetics of IL18R1 expression, we confirmed that IL18R1 is expressed on naïve cells at a level intermediate between Th1 and Th2 cells. We also found that T-bet is not absolutely required for IL18R1 expression and, while the expression of IL18R1 is decreased on T-bet deficient Th1 cells, it is not lost. T-bet deficient cells can signal in response to IL-18 *in vitro*. However, T-bet may be required for the stability of IL18R1 expression with time both *in vitro* and *in vivo*.

Examination of array data on cells skewed to different T helper cell lineages, suggests that IL18R1 and IL18RAP are not always co-regulated. As expected, expression of both IL18R1 and Il18RAP are high on Th1 cells and lower on Th2 cells. However, while IL18RAP expression is lower on Th17 cells than Th2, the expression of IL18R1 would seem to be intermediate between the two. IL18R1 expression is also seen on IL-17A producing cells in a mouse model of coeliac disease and the expression of IL18R1 is less affected by the loss of T-bet on this population than on IFN-γ producing cells. This might suggest that IL18R1 is regulated in two different ways in the two different cell types. However, we only performed the coeliac model once with small numbers of mice and we did not examine IL-17A production in the adoptive transfer model of colitis. Furthermore, we have very little data on the kinetics of IL18RAP expression at a protein level. Therefore, our conclusions must be treated with caution.

Basic next steps to this analysis would include the examination of IL18RAP expression on naïve cells and of both IL18R1 and IL18RAP expression on Th17 cells. It would also be helpful to examine the capacity of a pure population of naïve cells to signal in response to IL-18 as this would link in with studies suggesting a role for IL-18 signalling on naïve T cells in non Th1 like responses and asthma. The finding of other SNPs in the *IL18R1/IL18RAP* locus, in binding sites for GATA3 in Th1 and Th2 cells, and associated with asthma and white blood cell types, highlights the importance of this examination. The original GWAS for white blood cell types found that rs17027258, which is in strong LD with our hit-SNP rs3732123, is specifically associated with Eosinophil count.[321] Of note, our hit-SNPs, rs1465321 and rs373123, are in different LD blocks. The r^2

value between rs1465321 and rs12991737 is 0.0316, the r^2 between rs146532 and rs3732123 is 0.109 and the r^2 between rs12991737 and rs3732123 is 0.0544.

We must be careful in interpreting these GWAS in terms of IL-18 signalling as the locus also contains the gene *IL1RL1*, also named *IL33R* which is, in many respects the Th2 equivalent of *IL18R1*. IL-33 signalling is similar to IL-18 signalling but occurs in Th2 cells through a receptor complex of IL1RL1 and IL1RAP. Therefore, genomic architecture at this region can influence both a Th1 type and Th2 type response and little is known about whether that is independent or coordinated. It would seem that, at the *IL18R1/IL18RAP* locus different variants in different LD blocks can alter susceptibility to disease. Some genetic variation increases the Th1 response and thus increases risk of coeliac and Crohn's. A separate set of variations can increase Th2 type response and result in increased risk of asthma or a high eosinophil count. To build up a full picture of genomic architecture in the region, we would need 3C data on chromosomal conformation for the region in Th1, Th2 and naïve T cells.

A key transcription factor in IL18R1 regulation that is yet to be discussed is STAT4 which is needed for full expression of IL18R1 and IL18RAP. Yu *et al*[23] show that transient expression of STAT4 as a cell commits to a Th1 lineage generates a permissive epigenetic environment for IL18R1 and IL18RAP expression. They demonstrate that in mice, in response to IL-12, STAT4 becomes active, binds to a region upstream of the *IL18R1* TSS and promotes histone acetylation at both the promoter and first intron of *IL18R1*. They demonstrate that STAT4 binding and its effects peak at around 48 hours after addition of IL-12 but then fall over the next five days. They also demonstrate that STAT4 binding prevents repressive methylation of the DNA and the binding of DNA methyltransferases. It is possible that the role of maintaining such a permissive environment then falls, later, to T-bet. In mice, strong T-bet binding peaks are seen in the first, third and fourth introns. One of the peaks seen in the first intron is present in Th1 but not Treg cells suggesting it is Th1 specific and plays a key part in IL18R1 remodelling during Th1 lineage commitment. This might explain our results which suggest that IL18R1 expression is not absent but might be unstable over time in T-bet$^{-/-}$ cells. It is possible that the switch from temporarily activated expression of IL18R1 by STAT4 to stable IL18R1 expression maintained by T-bet is impaired in T-bet$^{-/-}$ mice. This might also explain why we found that IL18R1 expression on T-bet$^{-/-}$ cells was reduced but still present whereas Thieu

et al found no IL18R1 expression. The extent of IL18R1 expression in Tbet$^{-/-}$ cells could depend on the time since IL-12 signal was given which varied between our two experimental systems.[213] Of note, the first intron in the mouse *Il18r1* locus is not present in the UCSC gene schematic of human *IL18R1* (as shown in figures 4.12 and 6.5). However, this region is moderately conserved between mouse and human[204] and human transcripts both with and without this intron are listed in the Ensembl database. In mice, we find the most noticeable differential T-bet peak between Treg and Th1 cells in this first murine intron. In Treg cells, this peak corresponds to a peak of FoxP3 binding which may serve to limit IL18R1 expression in Tregs, although expression of IL18R1 on these cells has not been formally tested.

T-bet binding is seen upstream of the *IL18R1* TSS in human. The STAT4 binding site data that we used for our permutation testing had been obtained from sorted CD4$^+$, CD45RA$^+$ cells from human buffy coats which had been activated with αCD3 and αCD28 and stimulated in the presence of IL-12 so this data is a reasonable approximation to Th1 cells. These cells show STAT4 binding either side of the second exon at the human *IL18R1* locus, overlapping the two large T-bet peaks seen there. Our hit-SNP rs1465321 is found in a third T-bet peak just downstream and in a region of moderate homology as aligned in Ensembl. The regions either side of the second exon also contain histone modifications and a differential pattern of DNase hypersensitivity between Th1 and Th2 cells. The genomic marks of this region suggest it could be an enhancer region acting on the *IL18R1* promoter. GATA3 is present at the region in both Th1 and Th2 cells. In Th2 cells, it would seem to play an active part in repressing IL18R1 expression as retroviral transduction of constitutively expressed GATA3 into Th1 cells reduces expression of IL18R1.[204] A schematic of some of the dynamically changing features found by us and others at the IL18R1 locus in mouse and human can be seen in figure 6.5.

Another important region is the promoter of IL18RAP. Although we have focused less on this region, one of our hit-SNPs, rs1420106, is in this location in a binding site for T-bet and GATA3 in Th1 and Th2 cells and the SNP disrupts a GATA motif. As yet, we have not sufficiently established a system to test this for altered GATA3 binding and functional effect at this SNP but it would be a high priority once such assays are established. From a statistical genetics perspective, this SNP is interesting as it achieved a higher p-value in the recent coeliac

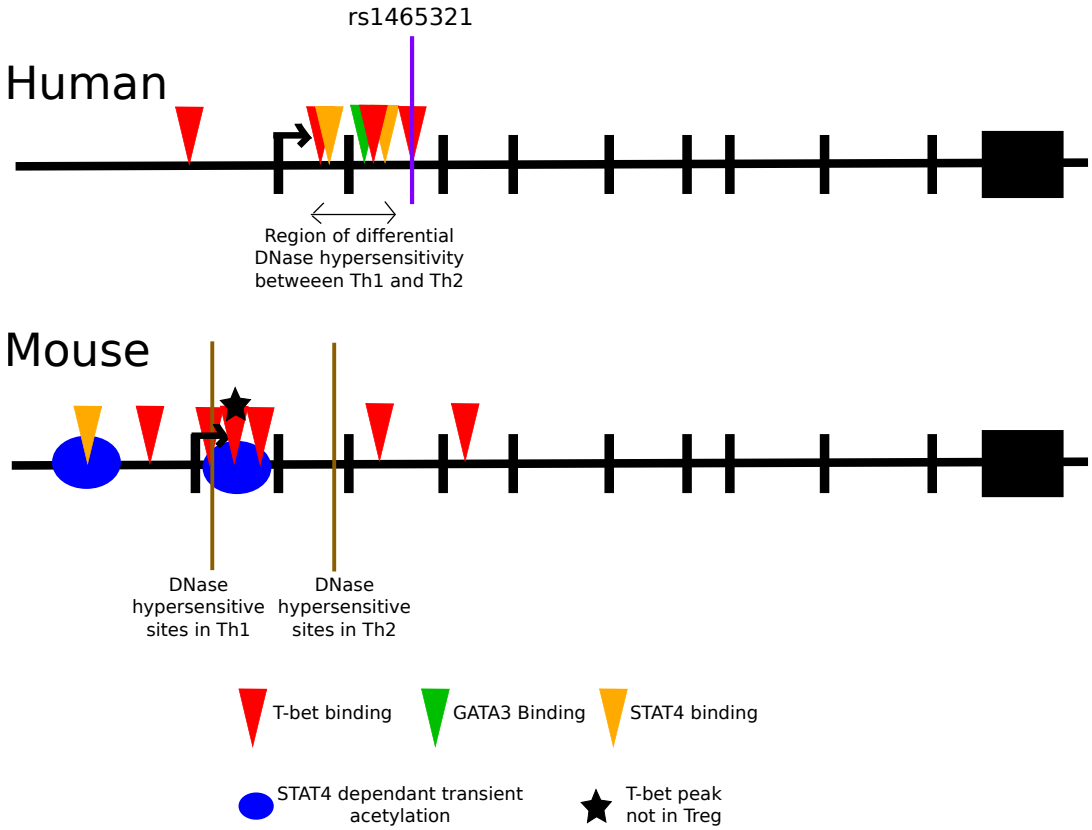

Figure 6.5: Features at the IL18R1 locus. - Some of the features at the IL18R1 locus in human (top) and mouse (bottom).

ImmunoChIP than rs1465321 and expression of IL18RAP is part of an eQTL in whole blood.[75,214]

6.4.2 Future Directions - The *IL18R1/IL18RAP* Locus

Although the data obtained so far gives some important indications about the kinetics of IL18R1 and IL18RAP expression, much of the work needs repeating before firm conclusions can be drawn. Although we can confidently conclude that lack of T-bet results in reduced, but not absent, IL18R1 expression, we would need to run the timecourse experiment again to confirm whether IL18R1 expression is unstable, peaking and then decreasing with time in T-bet deficient cells. As already mentioned, an investigation of the kinetics of IL18RAP expression would also be useful.

6.5 Role of T-bet in Model of Coeliac Disease

Because rs1465321 was in LD with a trait-associated SNP for coeliac disease, we examined the role of T-bet in a mouse model of coeliac disease. We found that, in agreement with the original paper and in line with the human disease, the model was milder than most models of IBD. As a result, it was not possible to observe any macroscopic effects of T-bet loss in the model. Furthermore, the model was inconsistent and due to time constraints, our group sizes were small. It may be that some patterns could be seen if we were to repeat the experiment. However, there are issues with the model. The disease develops over a long period of time in Rag1 deficient hosts. During this time, even the hosts given PBS develop some inflammation, as can be seen histologically. Dissecting out inflammation caused by the model versus that caused by other means is not easy.

One crucial finding of the model was the expression of IL18R1 on IL-17A$^+$ single producers and IL-17A$^+$, IFN-γ^+ double producers. IL18R1 expression seems to be dependent on T-bet only in those cells producing IFN-γ. Again, we cannot draw firm conclusions from the small numbers used but the examination of this is the context of coeliac and other disease models would be a useful next step.

6.6 Final Conclusions and Future Directions

In this study, we have attempted to combine data on disease-associated genetic variation with binding site data on transcription factors with known roles in disease and obtained in cells of appropriate lineage. We have then developed a medium throughput assay to test our *in silico* hits and tried to examine one of the DNA loci in a mouse model of disease. In doing so, we have attempted both to further investigate the mechanisms behind the regulation of a key immune gene and to evaluate different ways in which the growing datasets from genetic investigation can be examined with respect to detailed molecular mechanism.

Because of the novelty of the work presented it is difficult to draw broad-scale firm conclusions at this time. Our finding of differential binding at rs1465321, rs1006353 and rs11135484 suggest that out approach can find important hits for further investigation. However, there are limitations to the data obtained using our OligoFlow assay. Because we were simultaneously establishing the assay and collecting data on altered binding, there is variability between replicates for any one SNP, in terms of absolute MFI values obtained. Thus we cannot quantitatively analyse the data across replicates. As previously discussed, now that optimal parameters for the assay have been established, measures can be taken to run the assay so that absolute MFI values are comparable across time. Thus, replicate readings could be analysed quantitatively rather than simply looking for consistent trends as we have done. Furthermore, work is ongoing to establish a second assay where altered binding is analysed by ChIP followed by qPCR in cells that are heterozygous at the SNP of interest. If the results from this are consistent with those of the OligoFlow assay, it will validate the OligoFlow assay and provide data on the situation *in vivo*. As already mentioned our *in silico* work could also be improved by incorporation of the recently published 1000 genome data and use of some of the data and tools produced by the ENCODE project. (At the time of writing this thesis, certain aspects of ENCODE were still under embargo.)

Our work suggests that a study of the changing three dimensional structure of the *IL18R1/IL18RAP* locus in T helper cell lineage commitment would help better understand the regulation and role of this locus in different T helper cell subtypes. The possibility that IL18R1 is expressed on Th17 cells, and that its

expression is regulated differently in Th1 and Th17 cells, also remains to be examined. In terms of other hit-SNPs, it would be useful to expand the OligoFlow assay to other transcription factors. In addition, the role of T-bet at the locations around our hit-SNPs rs1006353 and rs11135484 remain to be investigated.

Although one of our hit-SNPs, rs1465321, was in the *IL18R1/IL18RAP* locus, our examination of IL18R1 expression in a mouse model of coeliac disease did not produce any significant results. This may be due to limited experimental group size. However, as discussed previously, there were issues with the model. A further direction might be to look at IL18R1 and IL18RAP expression with progression of disease in a different disease model. In this project, we briefly examined IL18R1 expression in a model of inflammatory bowel disease. The adoptive transfer model of inflammatory bowel disease is well established and gives a strong disease readout. Another next step would be to repeat and extend our work in this model.

References

[1] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome.** *Nature*, **409**(6822):860–921, February 2001. 19, 35, 36

[2] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigo, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. **The sequence of the human genome.** *Science*, **291**(5507):1304–1351, February 2001. 19, 35

[3] Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Bum-Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger, Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura A L Dillon,

Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Eric D Green, Peter J Good, Elise A Feingold, Bradley E Bernstein, Ewan Birney, Gregory E Crawford, Job Dekker, Laura Elnitski, Peggy J Farnham, Mark Gerstein, Morgan C Giddings, Thomas R Gingeras, Eric D Green, Roderic Guigó, Ross C Hardison, Timothy J Hubbard, Manolis Kellis, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, Michael Snyder, John A Stamatoyannopoulos, Scott A Tenenbaum, Zhiping Weng, Kevin P White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Morgan C Giddings, Bradley E Bernstein, Charles B Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D Ward, Robert C Altshuler, Matthew L Eaton, Manolis Kellis, Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakraborty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Brian A Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R Gingeras, Kate R Rosenbloom, Cricket A Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, W James Kent, Vanessa M Kirkup, Laurence R Meyer, Jeffrey C Long, Morgan Maddren, Brian J Raney, Terrence S Furey, Lingyun Song, Linda L Grasfeder, Paul G Giresi, and L... **An integrated encyclopedia of DNA elements in the human genome**. *Nature*, **489**(7414):57–74, September 2012. 19, 23, 25, 27, 34, 36, 37, 93, 96, 177, 179

[4] Shelley L Berger. **The complex language of chromatin regulation during transcription**. *Nature*, **447**(7143):407–412, May 2007. 20

[5] Michael J Ziller, Fabian Müller, Jing Liao, Yingying Zhang, Hongcang Gu, Christoph Bock, Patrick Boyle, Charles B Epstein, Bradley E Bernstein, Thomas Lengauer, Andreas Gnirke, and Alexander Meissner. **Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types**. *PLoS Genetics*, **7**(12):e1002389, December 2011. 20

[6] J Wang, J Zhuang, S Iyer, X Lin, T W Whitfield, M C Greven, B G Pierce, X Dong, A Kundaje, Y Cheng, O J Rando, E Birney, R M Myers, W S Noble, M Snyder, and Z Weng. **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors**. *Genome Research*, **22**(9):1798–1812, September 2012. 21, 24, 60

[7] Peter C J Janson, Malin E Winerdal, and Ola Winqvist. **At the crossroads of T helper lineage commitment—Epigenetics points the way**. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1790**(9):906–919, September 2009. 21

[8] Gang Wei, Lai Wei, Jinfang Zhu, Chongzhi Zang, Jane Hu-Li, Zhengju Yao, Kairong Cui, Yuka Kanno, Tae-Young Roh, Wendy T Watford, Dustin E Schones, Weiqun Peng, Hong-wei Sun, William E Paul, John J O'Shea, and Keji Zhao. **Global Mapping of H3K4me3 and H3K27me3 Reveals Specificity and Plasticity in Lineage Fate Determination of Differentiating CD4+ T Cells**. *Immunity*, **30**(1):155–167, January 2009. 21, 58

[9] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, and Keji Zhao. **Combinatorial patterns of histone acetylations and methylations in the human genome**. *Nature Genetics*, **40**(7):897–903, June 2008. 21, 96

[10] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. **High-Resolution Profiling of Histone Methylations in the Human Genome**. *Cell*, **129**(4):823–837, May 2007. 21, 70, 95, 179, 183

[11] Sara A Miller and Amy S Weinmann. **Molecular mechanisms by which T-bet regulates T-helper cell commitment**. *Immunological Reviews*, **238**(1):233–246, October 2010. 21

[12] Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, Rudolf Jaenisch, Alexandre Wagschal, Robert Feil, Stuart L Schreiber, and Eric S Lander. **A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells**. *Cell*, **125**(2):315–326, April 2006. 22

[13] Xianjun Dong, Melissa C Greven, Anshul Kundaje, Sarah Djebali, James B Brown, Chao Cheng, Tom R Gingeras, Mark Gerstein, Roderic Guigó, Ewan Birney, and Zhiping Weng. **Modeling gene expression using chromatin features in various cellular contexts.** *Genome biology*, **13**(9):R53, September 2012. 22

[14] Tony Kouzarides. **Chromatin Modifications and Their Function**. *Cell*, **128**(4):693–705, February 2007. 22

[15] Neil Macdonald, Julie P I Welburn, Martin E M Noble, Anhco Nguyen, Michael B Yaffe, David Clynes, Jonathan G Moggs, George Orphanides, Stuart Thomson, John W Edmunds, Alison L Clayton, Jane A Endicott, and Louis C Mahadevan. **Molecular Basis for the Recognition of Phosphorylated and Phosphoacetylated Histone H3 by 14-3-3**. *Molecular cell*, **20**(2):199–211, October 2005. 22

[16] Rushad Pavri, Bing Zhu, Guohong Li, Patrick Trojer, Subhrangsu Mandal, Ali Shilatifard, and Danny Reinberg. **Histone H2B Monoubiquitination Functions Cooperatively with FACT to Regulate Elongation by RNA Polymerase II**. *Cell*, **125**(4):703–717, May 2006. 22

[17] D Nathan. **Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications**. *Genes & Development*, **20**(8):966–976, April 2006. 22

[18] P G Giresi, J Kim, R M McDaniell, V R Iyer, and J D Lieb. **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin**. *Genome Research*, **17**(6):877–885, June 2007. 23

[19] B Ren, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T L Volkert, C J Wilson, S P Bell, and R A Young. **Genome-wide location and function of DNA binding proteins.** *Science*, **290**(5500):2306–2309, December 2000. 23

[20] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nature Methods*, **4**(8):651–657, August 2007. 23

[21] Shirley Pepke, Barbara Wold, and Ali Mortazavi. **Computation for ChIP-seq and RNA-seq studies**. *Nature Methods*, **6**(11s):S22–S32, November 2009. 23

[22] Elizabeth G Wilbanks and Marc T Facciotti. **Evaluation of algorithm performance in ChIP-seq peak detection.** *PLoS ONE*, **5**(7):e11471, 2010. 23, 116, 117

[23] Qing Yu, Vivian T Thieu, and Mark H Kaplan. **Stat4 limits DNA methyltransferase recruitment and DNA methylation of the IL-18Ralpha gene during Th1 differentiation.** *The EMBO journal*, **26**(8):2052–2060, April 2007. 24, 53, 201

[24] G Badis, M F Berger, A A Philippakis, S Talukder, A R Gehrke, S A Jaeger, E T Chan, G Metzler, A Vedenko, X Chen, H Kuznetsov, C F Wang, D Coburn, D E Newburger, Q Morris, T R Hughes, and M L Bulyk. **Diversity and Complexity in DNA Recognition by Transcription Factors**. *Science*, **324**(5935):1720–1723, June 2009. 24

[25] Seth Frietze, Rui Wang, Lijing Yao, Yu Gyoung Tak, Zhenqing Ye, Malaina Gaddis, Heather Witt, Peggy J Farnham, and Victor X Jin. **Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3.** *Genome biology*, **13**(9):R52, September 2012. 24

[26] J M Landolin, D S Johnson, N D Trinklein, S F Aldred, C Medina, H Shulha, Z Weng, and R M Myers. **Sequence features that drive human promoter function and tissue specificity**. *Genome Research*, **20**(7):890–898, July 2010. 24

[27] Troy W Whiteld, Jie Wang, Patrick J Collins, E Christopher Partridge, Shelley Force Aldred, Nathan D Trinklein, Richard M Myers, and Zhiping Weng. **Functional analysis of transcription factor binding sites in human promoters.** *Genome biology*, **13**(9):R50, September 2012. 24, 144, 191

[28] Tina Hummelshoj, Lars P Ryder, Hans O Madsen, Niels Odum, and Arne Svejgaard. **A functional polymorphism in the Eta-1 promoter is associated with allele specific binding to the transcription factor Sp1 and elevated gene expression**. *Molecular Immunology*, **43**(7):980–986, March 2006. 25

[29] L Larsson, L Rymo, and T Berglundh. **Sp1 binds to the G allele of the -1087 polymorphism in the IL-10 promoter and promotes IL-10 mRNA transcription and protein production**. *Genes and Immunity*, **11**(2):181–187, January 2010. 25

[30] Yi Cao, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J Sanchez, Maura H Parker, Kyle L MacQuarrie, Jerry Davison, Martin T Morgan, and Walter L Ruzzo. **Genome-wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming**. *Developmental Cell*, **18**(4):662–674, April 2010. 25

[31] M Kasowski, F Grubert, C Heffelfinger, M Hariharan, A Asabere, S M Waszak, L Habegger, J Rozowsky, M Shi, A E Urban, M Y Hong, K J Karczewski, W Huber, S M Weissman, M B Gerstein, J O Korbel, and M Snyder. **Variation in Transcription Factor Binding Among Humans**. *Science*, **328**(5975):232–235, April 2010. 25, 73

[32] Lisette Graae, Robert Karlsson, and Silvia Paddock. **Significant Association of Estrogen Receptor Binding Site Variation with Bipolar Disorder in Females**. *PLoS ONE*, **7**(2):e32304, February 2012. 25

[33] Bing Ge, Dmitry K Pokholok, Tony Kwan, Elin Grundberg, Lisanne Morcos, Dominique J Verlaan, Jennie Le, Vonda Koka, Kevin C L Lam, Vincent Gagné, Joana Dias, Rose Hoberman, Alexandre Montpetit, Marie-Michele Joly, Edward J Harvey, Daniel Sinnett, Patrick Beaulieu, Robert Hamon, Alexandru Graziani, Ken Dewar, Eef Harmsen, Jacek Majewski, Harald H H Göring, Anna K Naumova, Mathieu Blanchette, Kevin L Gunderson, and Tomi Pastinen. **Global patterns of cis variation in human cells revealed by high-density allelic expression analysis**. *Nature Genetics*, **41**(11):1216–1222, October 2009. 25

[34] Nathaniel D Maynard, Jing Chen, Rhona K Stuart, Jian-Bing Fan, and Bing Ren. **Genome-wide mapping of allele-specific protein-DNA interactions in human cells**. *Nature Methods*, March 2008. 25

[35] H P Müeller-Storm, J M Sogo, and W Schaffner. **An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge.** *Cell*, **58**(4):767–777, August 1989. 26

[36] J Dekker. **Capturing Chromosome Conformation**. *Science*, **295**(5558):1306–1311, February 2002. 26

[37] Marieke Simonis, Jurgen Kooren, and Wouter de Laat. **An evaluation of 3C-based methods to capture DNA interactions.** *Nature Methods*, **4**(11):895–901, November 2007. 26

[38] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C)**. *Nature Genetics*, **38**(11):1348–1354, October 2006. 26

[39] J Dostie, T A Richmond, R A Arnaout, R R Selzer, W L Lee, T A Honan, E D Rubio, A Krumm, J Lamb, C Nusbaum, R D Green, and J Dekker. **Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements**. *Genome Research*, **16**(10):1299–1309, October 2006. 26

[40] E Lieberman-Aiden, N L van Berkum, L Williams, M Imakaev, T Ragoczy, A Telling, I Amit, B R Lajoie, P J Sabo, M O Dorschner, R Sandstrom, B Bernstein, M A Bender, M Groudine, A Gnirke, J Stamatoyannopoulos, L A Mirny, E S Lander, and J Dekker. **Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**. *Science*, **326**(5950):289–293, October 2009. 27

[41] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. **An oestrogen-receptor-α-bound human chromatin interactome**. *Nature*, **462**(7269):58–64, November 2009. 27

[42] Karen M Lower, Jim R Hughes, Marco De Gobbi, Shirley Henderson, Vip Viprakasit, Chris Fisher, Anne Goriely, Helena Ayyub, Jackie Sloane-Stanley, Douglas Vernimmen,

CORDELIA LANGFORD, DAVID GARRICK, RICHARD J GIBBONS, AND DOUGLAS R HIGGS. **Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition.** *Proceedings of the National Academy of Sciences of the United States of America*, **106**(51):21771–21776, December 2009. 27

[43] CHARALAMPOS G SPILIANAKIS, MARIA D LALIOTI, TERRENCE TOWN, GAP RYOL LEE, AND RICHARD A FLAVELL. **Interchromosomal associations between alternatively expressed loci**. *Nature*, **435**(7042):637–645, May 2005. 27, 48

[44] YIN SHEN, FENG YUE, DAVID F MCCLEARY, ZHEN YE, LEE EDSALL, SAMANTHA KUAN, ULRICH WAGNER, JESSE DIXON, LEONARD LEE, VICTOR V LOBANENKOV, AND BING REN. **A map of the cis-regulatory sequences in the mouse genome**. *Nature*, July 2012. 28

[45] S FISHER. **Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity**. *Science*, **312**(5771):276–279, April 2006. 28

[46] A S DIMAS, S DEUTSCH, B E STRANGER, S B MONTGOMERY, C BOREL, H ATTAR-COHEN, C INGLE, C BEAZLEY, M G ARCELUS, M SEKOWSKA, M GAGNEBIN, J NISBETT, P DELOUKAS, E T DERMITZAKIS, AND S E ANTONARAKIS. **Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner**. *Science*, **325**(5945):1246–1250, September 2009. 28, 139

[47] HYUNSEOK PETER KANG, ALEX A MORGAN, RONG CHEN, ERIC E SCHADT, AND ATUL J BUTTE. **Coanalysis of GWAS with eQTLs reveals disease-tissue associations.** *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, **2012**:35–41, 2012. 28

[48] MARY MUERS. **Gene expression: Transcriptome to proteome and back to genome**. *Nature Reviews Genetics*, **12**(8):518–518, June 2011. 28

[49] BJÖRN SCHWANHÄUSSER, DOROTHEA BUSSE, NA LI, GUNNAR DITTMAR, JOHANNES SCHUCHHARDT, JANA WOLF, WEI CHEN, AND MATTHIAS SELBACH. **Global quantification of mammalian gene expression control**. *Nature*, **473**(7347):337–342, May 2011. 28

[50] MATTHEW HALVORSEN, JOSHUA S MARTIN, SAM BROADAWAY, AND ALAIN LAEDERACH. **Disease-Associated Mutations That Alter the RNA Structural Ensemble**. *PLoS Genetics*, **6**(8):e1001074, August 2010. 28

[51] K L CARTER, E CAHIR-MCFARLAND, AND E KIEFF. **Epstein-Barr Virus-Induced Changes in B-Lymphocyte Gene Expression**. *Journal of Virology*, **76**(20):10427–10436, October 2002. 29

[52] M CALISKAN, D A CUSANOVICH, C OBER, AND Y GILAD. **The effects of EBV transformation on gene expression levels and methylation profiles**. *Human Molecular Genetics*, **20**(8):1643–1652, March 2011. 29

[53] K BULLAUGHEY, C I CHAVARRIA, G COOP, AND Y GILAD. **Expression quantitative trait loci detected in cell lines are often present in primary tissues**. *Human Molecular Genetics*, **18**(22):4296–4303, October 2009. 29

[54] SHAI S SHEN-ORR, ROBERT TIBSHIRANI, PURVESH KHATRI, DALE L BODIAN, FRANK STAEDTLER, NICHOLAS M PERRY, TREVOR HASTIE, MINNIE M SARWAL, MARK M DAVIS, AND ATUL J BUTTE. **Cell type–specific gene expression differences in complex tissues**. *Nature Methods*, **7**(4):287–289, March 2010. 29

[55] A E PASQUINELLI, B J REINHART, F SLACK, M Q MARTINDALE, M I KURODA, B MALLER, D C HAYWARD, E E BALL, B DEGNAN, P MÜLLER, J SPRING, A SRINIVASAN, M FISHMAN, J FINNERTY, J CORBO, M LEVINE, P LEAHY, E DAVIDSON, AND G RUVKUN. **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA.** *Nature*, **408**(6808):86–89, November 2000. 29

[56] LI-FAN LU, MARK P BOLDIN, ASHUTOSH CHAUDHRY, LING-LI LIN, KONSTANTIN D TAGANOV, TOSHIKATSU HANADA, AKIHIKO YOSHIMURA, DAVID BALTIMORE, AND ALEXANDER Y RUDENSKY. **Function of miR-146a in Controlling Treg Cell-Mediated Regulation of Th1 Responses**. *Cell*, **142**(6):914–929, September 2010. 29

[57] ÁNGELES MENCÍA, SILVIA MODAMIO-HØYBJØR, NICK REDSHAW, MATÍAS MORÍN, FERNANDO MAYO-MERINO, LETICIA OLAVARRIETA, LUIS A AGUIRRE, IGNACIO DEL CASTILLO, KAREN P STEEL, TAMAS DALMAY, FELIPE MORENO, AND MIGUEL ÁNGEL MORENO-PELAYO. **Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss**. *Nature Genetics*, **41**(5):609–613, April 2009. 29

[58] LOÏC DE PONTUAL, EVELYN YAO, PATRICK CALLIER, LAURENCE FAIVRE, VALÉRIE DROUIN, SANDRA CARIOU, ARIE VAN HAERINGEN, DAVID GENEVIÈVE, ALICE GOLDENBERG, MYRIAM OUFADEM, SYLVIE MANOUVRIER, ARNOLD MUNNICH, JOANA ALVES VIDIGAL, MICHEL VEKEMANS, STANISLAS LYONNET, ALEXANDRA HENRION-CAUDE, ANDREA VENTURA, AND JEANNE AMIEL. **Germline deletion of the miR-17âĹij92 cluster causes skeletal and growth defects in humans**. *Nature Genetics*, **43**(10):1026–1030, September 2011. 30

[59] YI W KONG, DAVID FERLAND-MCCOLLOUGH, THOMAS J JACKSON, AND MARTIN BUSHELL. **microRNAs in cancer management.** *The lancet oncology*, **13**(6):e249–58, June 2012. 30

[60] JOSHUA T MENDELL AND ERIC N OLSON. **MicroRNAs in Stress Signaling and Human Disease**. *Cell*, **148**(6):1172–1187, March 2012. 30

[61] SIMONA ROSSI, ARISTOTELIS TSIRIGOS, ANTONIO AMOROSO, NICOLETTA MASCELLANI, ISIDORE RIGOUTSOS, GEORGE A CALIN, AND STEFANO VOLINIA. **OMiR: Identification of associations between OMIM diseases and microRNAs**. *Genomics*, **97**(2):71–76, February 2011.

[62] KRIS RICHARDSON, CHAO-QIANG LAI, LAURENCE D PARNELL, YU-CHI LEE, AND JOSE M ORDOVAS. **A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS.** *BMC genomics*, **12**:504, 2011. 30

[63] T DERRIEN, R JOHNSON, G BUSSOTTI, A TANZER, S DJEBALI, H TILGNER, G GUERNEC, D MARTIN, A MERKEL, D G KNOWLES, J LAGARDE, L VEERAVALLI, X RUAN, Y RUAN, T LASSMANN, P CARNINCI, J B BROWN, L LIPOVICH, J M GONZALEZ, M THOMAS, C A DAVIS, R SHIEKHATTAR, T R GINGERAS, T J HUBBARD, C NOTREDAME, J HARROW, AND R GUIGO. **The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression**. *Genome Research*, **22**(9):1775–1789, September 2012. 30

[64] S P COLLIER, P L COLLINS, C L WILLIAMS, M R BOOTHBY, AND T M AUNE. **Cutting Edge: Influence of Tmevpg1, a Long Intergenic Noncoding RNA, on the Expression of Ifng by Th1 Cells**. *The Journal of Immunology*, July 2012. 30

[65] JOEL N HIRSCHHORN AND MARK J DALY. **Genome-wide association studies for common diseases and complex traits**. *Nature Reviews Genetics*, **6**(2):95–108, February 2005. 30, 31

[66] S J DOCHERTY, O S P DAVIS, Y KOVAS, E L MEABURN, P S DALE, S A PETRILL, L C SCHALKWYK, AND R PLOMIN. **A genome-wide association study identifies multiple loci associated with mathematics ability and disability**. *Genes, Brain and Behavior*, **9**(2):234–247, March 2010. 31

[67] Herbert W Virgin and John A Todd. **Metagenomics and Personalized Medicine**. *Cell*, **147**(1):44–56, September 2011. 31, 62, 63, 64

[68] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo Y K Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, Yong Cheng, Michael J Clark, Hogune Im, Lukas Habegger, Suganthi Balasubramanian, Maeve O'Huallachain, Joel T Dudley, Sara Hillenmeyer, Rajini Haraksingh, Donald Sharon, Ghia Euskirchen, Phil Lacroute, Keith Bettinger, Alan P Boyle, Maya Kasowski, Fabian Grubert, Scott Seki, Marco Garcia, Michelle Whirl-Carrillo, Mercedes Gallardo, Maria A Blasco, Peter L Greenberg, Phyllis Snyder, Teri E Klein, Russ B Altman, Atul J Butte, Euan A Ashley, Mark Gerstein, Kari C Nadeau, Hua Tang, and Michael Snyder. **Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes**. *Cell*, **148**(6):1293–1307, March 2012. 31

[69] Hakon Hakonarson and Struan F A Grant. **Planning a genome-wide association study: Points to consider**. *Annals of Medicine*, **43**(6):451–460, September 2011. 31

[70] R J Klein. **Complement Factor H Polymorphism in Age-Related Macular Degeneration**. *Science*, **308**(5720):385–389, April 2005. 31

[71] Jochen Hampe, Andre Franke, Philip Rosenstiel, Andreas Till, Markus Teuber, Klaus Huse, Mario Albrecht, Gabriele Mayr, Francisco M De La Vega, Jason Briggs, Simone Günther, Natalie J Prescott, Clive M Onnie, Robert Häsler, Bence Sipos, Ulrich R Fölsch, Thomas Lengauer, Matthias Platzer, Christopher G Mathew, Michael Krawczak, and Stefan Schreiber. **A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1**. *Nature Genetics*, **39**(2):207–211, December 2006. 31, 63

[72] WG Hill and A Robertson. **Linkage disequilibrium in finite populations**. *TAG Theoretical and Applied Genetics*, **38**(6):226–231, 1968. 32

[73] The International HapMap Consortium. **A haplotype map of the human genome**. *Nature*, **437**(7063):1299–1320, October 2005. 32

[74] David M Altshuler, Richard A Gibbs, Leena Peltonen, David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Leena Peltonen, Emmanouil Dermitzakis, Penelope E Bonnen, David M Altshuler, Richard A Gibbs, Paul I W de Bakker, Panos Deloukas, Stacey B Gabriel, Rhian Gwilliam, Sarah Hunt, Michael Inouye, Xiaoming Jia, Aarno Palotie, Melissa Parkin, Pamela Whittaker, Fuli Yu, Kyle Chang, Alicia Hawes, Lora R Lewis, Yanru Ren, David Wheeler, Richard A Gibbs, Donna Marie Muzny, Chris Barnes, Katayoon Darvishi, Matthew Hurles, Joshua M Korn, Kati Kristiansson, Charles Lee, Steven A McCarroll, James Nemesh, Emmanouil Dermitzakis, Alon Keinan, Stephen B Montgomery, Samuela Pollack, Alkes L Price, Nicole Soranzo, Penelope E Bonnen, Richard A Gibbs, Claudia Gonzaga-Jauregui, Alon Keinan, Alkes L Price, Fuli Yu, Verneri Anttila, Wendy Brodeur, Mark J Daly, Stephen Leslie, Gil McVean, Loukas Moutsianas, Huy Nguyen, Stephen F Schaffner, Qingrun Zhang, Mohammed J R Ghori, Ralph McGinnis, William McLaren, Samuela Pollack, Alkes L Price, Stephen F Schaffner, Fumihiko Takeuchi, Sharon R Grossman, Ilya Shlyakhter, Elizabeth B Hostetter, Pardis C Sabeti, Clement A Adebamowo, Morris W Foster, Deborah R Gordon, Julio Licinio, Maria Cristina Manca, Patricia A Marshall, Ichiro Matsuda, Duncan Ngare, Vivian Ota Wang, Deepa Reddy, Charles N Rotimi, Charmaine D Royal, Richard R Sharp, Changqing Zeng, Lisa D Brooks, and Jean E McEwen. **Integrating common and rare genetic variation in diverse human populations**. *Nature*, **467**(7311):52–58, September 2010. 32, 69

[75] Gosia Trynka, Karen A Hunt, Nicholas A Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, Gemma Castillejo, Emilio G de la Concha, Rodrigo Coutinho de Almeida, Kerith-Rae M Dias, Cleo C van Diemen, Patrick C A Dubois, Richard H Duerr, Sarah Edkins, Lude Franke, Karin Fransen, Javier Gutierrez, Graham A R Heap, Barbara Hrdlickova, Sarah Hunt, Leticia Plaza Izurieta, Valentina Izzo, Leo A B Joosten, Cordelia Langford, Maria Cristina Mazzilli, Charles A Mein, Vandana Midah, Mitja Mitrovic, Barbara Mora, Marinita Morelli, Sarah Nutland, Concepción Núñez, Suna Onengut-Gumuscu, Kerra Pearce, Mathieu Platteel, Isabel Polanco, Simon Potter, Carmen Ribes-Koninckx, Isis Ricaño-Ponce, Stephen S Rich, Anna Rybak, José Luis Santiago, Sabyasachi Senapati, Ajit Sood, Hania Szajewska, Riccardo Troncone, Jezabel Varadé, Chris Wallace, Victorien M Wolters, Alexandra Zhernakova, B K Thelma, Bozena Cukrowska, Elena Urcelay, Jose Ramon Bilbao, M Luisa Mearin, Donatella Barisani, Jeffrey C Barrett, Vincent Plagnol, Panos Deloukas, Cisca Wijmenga, and David A van Heel. **Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease**. *Nature Genetics*, **43**(12):1193–1201, November 2011. 33, 199, 204

[76] Olivier Harismendy, Dimple Notani, Xiaoyuan Song, Nazli G Rahim, Bogdan Tanasa, Nathaniel Heintzman, Bing Ren, Xiang-Dong Fu, Eric J Topol, Michael G Rosenfeld, and Kelly A Frazer. **9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response**. *Nature*, **470**(7333):264–268, February 2011. 33

[77] Pauline C Ng, Samuel Levy, Jiaqi Huang, Timothy B Stockwell, Brian P Walenz, Kelvin Li, Nelson Axelrod, Dana A Busam, Robert L Strausberg, and J Craig Venter. **Genetic variation in an individual human exome.** *PLoS Genetics*, **4**(8):e1000160, 2008. 33

[78] Hironori Ueda, Joanna M M Howson, Laura Esposito, Joanne Heward, Hywel Snook, Giselle Chamberlain, Daniel B Rainbow, Kara M D Hunter, Annabel N Smith, Gianfranco Di Genova, Mathias H Herr, Ingrid Dahlman, Felicity Payne, Deborah Smyth, Christopher Lowe, Rebecca C J Twells, Sarah Howlett, Barry Healy, Sarah Nutland, Helen E Rance, Vin Everett, Luc J Smink, Alex C Lam, Heather J Cordell, Neil M Walker, Cristina Bordin, John Hulme, Costantino Motzo, Francesco Cucca, J Fred Hess, Michael L Metzker, Jane Rogers, Simon Gregory, Amit Allahabadia, Ratnasingam Nithiyananthan, Eva Tuomilehto-Wolf, Jaakko Tuomilehto, Polly Bingley, Kathleen M Gillespie, Dag E Undlien, Kjersti S Rønningen, Cristian Guja, Constantin Ionescu-Tîrgovişte, David A Savage, A Peter Maxwell, Dennis J Carson, Chris C Patterson, Jayne A Franklyn, David G Clayton, Laurence B Peterson, Linda S Wicker, John A Todd, and Stephen C L Gough. **Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease.** *Nature*, **423**(6939):506–511, May 2003. 33

[79] David F Burke, Catherine L Worth, Eva-Maria Priego, Tammy Cheng, Luc J Smink, John A Todd, and Tom L Blundell. **Genome bioinformatic analysis of nonsynonymous SNPs**. *BMC Bioinformatics*, **8**(1):301, 2007. 34

[80] N L Sim, P Kumar, J Hu, S Henikoff, G Schneider, and P C Ng. **SIFT web server: predicting effects of amino acid substitutions on proteins**. *Nucleic Acids Research*, **40**(W1):W452–W457, June 2012. 34

[81] Rong Chen, Eugene V Davydov, Marina Sirota, and Atul J Butte. **Non-Synonymous and Synonymous Coding SNPs Show Similar Likelihood and Effect Size of Human Disease Association**. *PLoS ONE*, **5**(10):e13574, October 2010. 34

[82] Greg Gibson. **Rare and common variants: twenty arguments**. *Nature Reviews Genetics*, **13**(2):135–145, January 2012. 34

[83] Richard M Durbin, David L Altshuler, Richard M Durbin, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Francis S Collins, Francisco M De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Bartha M Knoppers, Eric S Lander, Hans Lehrach, Elaine R Mardis, Gil A McVean, Debbie A Nickerson, Leena Peltonen, Alan J Schafer, Stephen T Sherry, Jun Wang, Richard K Wilson, Richard A Gibbs, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jun Wang, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Eric S Lander, David L Altshuler, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J Fennell, Stacey B Gabriel, David B Jaffe, Erica Shefler, Carrie L Sougnez, David R Bentley, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer Stone, Kevin J McKernan, Gina L Costa, Jeffry K Ichikawa, Clarence C Lee, Ralf Sudbrak, Hans Lehrach, Tatiana A Borodina, Andreas Dahl, Alexey N Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V Soldatov, Bernd Timmermann, Marius Tolzmann, Michael Egholm, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, Elaine R Mardis, Richard K Wilson, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, Richard M Durbin, John Burton, David M Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Richard A Gibbs, David Wheeler, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Jun Wang, Xiaodong Fang, Xiaosen Guo, Ruiqiang Li, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Guoqing Li, Jian Wang, Huanming Yang, Gabor T Marth, Erik P Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R Quinlan, Chip Stewart, Michael P Stromberg, Alistair N Ward, Jiantao Wu, Charles Lee, Ryan E Mills, Xinghua Shi, Mark J Daly, Mark A DePristo, David L Altshuler, Aaron D Ball, Eric Banks, Toby Bloom, Brian L Browning, Kristian Cibulskis, Tim J Fennell, Kiran V Garimella, Sharon R Grossman, Robert E Handsaker, Matt Hanna, Chris Hartl, David B Jaffe, Andrew M Kernytsky, Joshua M Korn, Heng Li, Jared R Maguire, Steven A McCarroll, Aaron McKenna, James C Nemesh, Anthony A Philippakis, Ryan E Poplin, Alkes Price, Manuel A Rivas, Pardis C Sabeti, Stephen F Schaffner, Erica Shefler, Ilya A Shlyakhter, David N Cooper, Edward V Ball, Matthew Mort, Andrew D Phillips, Peter D Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C Yoon, Carlos D Bustamante, Andrew G Clark, Adam Boyko, Jeremiah Degenhardt, Simon Gravel, Ryan N Gutenkunst, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Paul Flicek, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M McLaren, Rajesh Radhakrishnan, Richard E Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O Korbel, Adrian M Stütz, Sean Humphray, Markus Bauer, R Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Aravinda Chakravarti, Kai Ye, Francisco M De La Vega, Yutao Fu, Fiona C L Hyland, Jonathan M Manning, Stephen F McLaughlin, and Heather... Peckham. **A map of human genome variation from population-scale sequencing**. *Nature*, **467**(7319):1061–1073, October 2010. 34

[84] P Vineis and N E Pearce. **Genome-wide association studies may be misinterpreted: genes versus heritability**. *Carcinogenesis*, **32**(9):1295–1298, September 2011. 34

[85] S W Guo. **Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting**. *American journal of human genetics*, **63**(1):252–258, 1998. 35

[86] S W Guo. **Familial aggregation of environmental risk factors and familial aggregation of disease.** *American journal of epidemiology*, **151**(11):1121–1131, June 2000. 35

[87] Chirag J Patel, Jayanta Bhattacharya, and Atul J Butte. **An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus**. *PLoS ONE*, **5**(5):e10746, May 2010. 35

[88] Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. **Epigenome-wide association studies for common human diseases**. *Nature Reviews Genetics*, **12**(8):529–541, July 2011. 35

[89] Amy Strange, Francesca Capon, Chris C A Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Céline Bellenguez, Judith G M Bergboer, Jenefer M Blackwell, Elvira Bramon, Suzannah J Bumpstead, Juan P Casas, Michael J Cork, Aiden Corvin, Panos Deloukas, Alexander Dilthey, Audrey Duncanson, Sarah Edkins, Xavier Estivill, Oliver Fitzgerald, Colin Freeman, Emiliano Giardina, Emma Gray, Angelika Hofer, Ulrike Hüffmeier, Sarah E Hunt, Alan D Irvine, Janusz Jankowski, Brian Kirby, Cordelia Langford, Jesús Lascorz, Joyce Leman, Stephen Leslie, Lotus Mallbris, Hugh S Markus, Christopher G Mathew, W H Irwin McLean, Ross McManus, Rotraut Mössner, Loukas Moutsianas, Åsa T Naluai, Frank O Nestle, Giuseppe Novelli, Alexandros Onoufriadis, Colin N A Palmer, Carlo Perricone, Matti Pirinen, Robert Plomin, Simon C Potter, Ramon M Pujol, Anna Rautanen, Eva Riveira-Munoz, Anthony W Ryan, Wolfgang Salmhofer, Lena Samuelsson, Stephen J Sawcer, Joost Schalkwijk, Catherine H Smith, Mona Ståhle, Zhan Su, Rachid Tazi-Ahnini, Heiko Traupe, Ananth C Viswanathan, Richard B Warren, Wolfgang Weger, Katarina Wolk, Nicholas Wood, Jane Worthington, Helen S Young, Patrick L J M Zeeuwen, Adrian Hayday, A David Burden, Christopher E M Griffiths, Juha Kere, André Reis, Gilean McVean, David M Evans, Matthew A Brown, Jonathan N Barker, Leena Peltonen, Peter Donnelly, and Richard C Trembath. **A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1**. *Nature Genetics*, **42**(11):985–990, October 2010. 35

[90] Claude Bouchard. **Gene-environment interactions in the etiology of obesity: defining the fundamentals.** *Obesity (Silver Spring, Md.)*, **16 Suppl 3**:S5–S10, December 2008. 35

[91] Boyd A Swinburn, Gary Sacks, Kevin D Hall, Klim McPherson, Diane T Finegood, Marjory L Moodie, and Steven L Gortmaker. **The global obesity pandemic: shaped by global drivers and local environments.** *Lancet*, **378**(9793):804–814, August 2011. 35

[92] Chen Wu, Peter Kraft, Kan Zhai, Jiang Chang, Zhaoming Wang, Yun Li, Zhibin Hu, Zhonghu He, Weihua Jia, Christian C Abnet, Liming Liang, Nan Hu, Xiaoping Miao, Yifeng Zhou, Zhihua Liu, Qimin Zhan, Yu Liu, Yan Qiao, Yuling Zhou, Guangfu Jin, Chuanhai Guo, Changdong Lu, Haijun Yang, Jianhua Fu, Dianke Yu, Neal D Freedman, Ti Ding, Wen Tan, Alisa M Goldstein, Tangchun Wu, Hongbing Shen, Yang Ke, Yixin Zeng, Stephen J Chanock,

PHILIP R TAYLOR, AND DONGXIN LIN. **Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions**. *Nature Genetics*, September 2012. 35

[93] E PENNISI. **HUMAN GENOME:Academic Sequencers Challenge Celera in a Sprint to the Finish**. *Science*, **283**(5409):1822–1823, March 1999. 35

[94] THOMAS P LARSSON, CHRISTIAN G MURRAY, TOBIAS HILL, ROBERT FREDRIKSSON, AND HELGI B SCHIÖTH. **Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery**. *FEBS Letters*, **579**(3):690–698, January 2005. 36

[95] W J KENT, C W SUGNET, T S FUREY, K M ROSKIN, T H PRINGLE, A M ZAHLER, AND A D HAUSSLER. **The Human Genome Browser at UCSC**. *Genome Research*, **12**(6):996–1006, May 2002. 36

[96] ROBERT C GENTLEMAN, VINCENT J CAREY, DOUGLAS M BATES, BEN BOLSTAD, MARCEL DETTLING, SANDRINE DUDOIT, BYRON ELLIS, LAURENT GAUTIER, YONGCHAO GE, JEFF GENTRY, KURT HORNIK, TORSTEN HOTHORN, WOLFGANG HUBER, STEFANO IACUS, RAFAEL IRIZARRY, FRIEDRICH LEISCH, CHENG LI, MARTIN MAECHLER, ANTHONY J ROSSINI, GUNTHER SAWITZKI, COLIN SMITH, GORDON SMYTH, LUKE TIERNEY, JEAN Y H YANG, AND JIANHUA ZHANG. **Bioconductor: open software development for computational biology and bioinformatics.** *Genome biology*, **5**(10):R80, 2004. 36, 69

[97] R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0. 36, 69

[98] JEREMY GOECKS, ANTON NEKRUTENKO, JAMES TAYLOR, AND GALAXY TEAM. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome biology*, **11**(8):R86, 2010. 36, 70, 179

[99] B GIARDINE. **Galaxy: A platform for interactive large-scale genome analysis**. *Genome Research*, **15**(10):1451–1455, September 2005.

[100] DANIEL BLANKENBERG, GREGORY VON KUSTER, NATHANIEL CORAOR, GURUPRASAD ANANDA, ROSS LAZARUS, MARY MANGAN, ANTON NEKRUTENKO, AND JAMES TAYLOR. **Galaxy: a web-based genome analysis tool for experimentalists.** *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 19**:Unit 19.10.1–21, January 2010. 36, 70, 179

[101] CYDNEY B NIELSEN, MICHAEL CANTOR, INNA DUBCHAK, DAVID GORDON, AND TING WANG. **Visualizing genomes: techniques and challenges**. *Nature Methods*, **7**(3s):S5–S15, February 2010. 36

[102] THE ENCODE PROJECT CONSORTIUM. **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science*, **306**(5696):636–640, October 2004. 36

[103] EWAN BIRNEY, JOHN A STAMATOYANNOPOULOS, ANINDYA DUTTA, RODERIC GUIGÓ, THOMAS R GINGERAS, ELLIOTT H MARGULIES, ZHIPING WENG, MICHAEL SNYDER, EMMANOUIL T DERMITZAKIS, JOHN A STAMATOYANNOPOULOS, ROBERT E THURMAN, MICHAEL S KUEHN, CHRISTOPHER M TAYLOR, SHANE NEPH, CHRISTOPH M KOCH, SAURABH ASTHANA, ANKIT MALHOTRA, IVAN ADZHUBEI, JASON A GREENBAUM, ROBERT M ANDREWS, PAUL FLICEK, PATRICK J BOYLE, HUA CAO, NIGEL P CARTER, GAYLE K CLELLAND, SEAN DAVIS, NATHAN DAY, PAWANDEEP DHAMI, SHANE C DILLON, MICHAEL O DORSCHNER, HEIKE FIEGLER, PAUL G GIRESI, JEFF GOLDY, MICHAEL HAWRYLYCZ, ANDREW HAYDOCK, RICHARD HUMBERT, KEITH D JAMES, BRETT E JOHNSON, ERICKA M JOHNSON, TRISTAN T FRUM, ELIZABETH R ROSENZWEIG, NEERJA KARNANI, KIRSTEN LEE, GREGORY C LEFEBVRE, PATRICK A NAVAS, FIDENCIO NERI, STEPHEN C J PARKER, PETER J SABO, RICHARD SANDSTROM, ANTHONY SHAFER, DAVID VETRIE, MOLLY WEAVER, SARAH WILCOX, MAN YU1, FRANCIS S COLLINS, JOB DEKKER, JASON D LIEB, THOMAS D TULLIUS, GREGORY E CRAWFORD, SHAMIL SUNYAEV, WILLIAM S NOBLE, IAN DUNHAM, ANINDYA DUTTA, RODERIC GUIGÓ, FRANCE DENOEUD, ALEXANDRE REYMOND, PHILIPP KAPRANOV, JOEL ROZOWSKY, DEYOU ZHENG, ROBERT CASTELO, ADAM FRANKISH, JENNIFER HARROW, SRINKA GHOSH, ALBIN SANDELIN, IVO L HOFACKER, ROBERT BAERTSCH, DAMIAN KEEFE, PAUL FLICEK, SUJIT DIKE, JILL CHENG, HEATHER A HIRSCH, EDWARD A SEKINGER, JULIEN LAGARDE, JOSEP F ABRIL, ATIF SHAHAB, CHRISTOPH FLAMM, CLAUDIA FRIED, JÖRG HACKERMÜLLER, JANA HERTEL, MANJA LINDEMEYER, KRISTIN MISSAL, ANDREA TANZER, STEFAN WASHIETL, JAN KORBEL, OLOF EMANUELSSON, JAKOB S PEDERSEN, NANCY HOLROYD, RUTH TAYLOR, DAVID SWARBRECK, NICHOLAS MATTHEWS, MARK C DICKSON, DARYL J THOMAS, MATTHEW T WEIRAUCH, JAMES GILBERT, JORG DRENKOW, IAN BELL, XIAODONG ZHAO, K G SRINIVASAN, WING-KIN SUNG, HONG SAIN OOI, KUO PING CHIU, SYLVAIN FOISSAC, TYLER ALIOTO, MICHAEL BRENT, LIOR PACHTER, MICHAEL L TRESS, ALFONSO VALENCIA, SIEW WOH CHOO, CHIOU YU CHOO, CATHERINE UCLA, CAROLINE MANZANO, CARINE WYSS, EVELYN CHEUNG, TAANE G CLARK, JAMES B BROWN, MADHAVAN GANESH, SANDEEP PATEL, HARI TAMMANA, JACQUELINE CHRAST, CHARLOTTE N HENRICHSEN, CHIKATOSHI KAI, JUN KAWAI, UGRAPPA NAGALAKSHMI, JIAQIAN WU, ZHENG LIAN, JIN LIAN, PETER NEWBURGER, XUEQING ZHANG, PETER BICKEL, JOHN S MATTICK, PIERO CARNINCI, YOSHIHIDE HAYASHIZAKI, SHERMAN WEISSMAN, EMMANOUIL T DERMITZAKIS, ELLIOTT H MARGULIES, TIM HUBBARD, RICHARD M MYERS, JANE ROGERS, PETER F STADLER, TODD M LOWE, CHIA-LIN WEI, YIJUN RUAN, MICHAEL SNYDER, EWAN BIRNEY, KEVIN STRUHL, MARK GERSTEIN, STYLIANOS E ANTONARAKIS, THOMAS R GINGERAS, JAMES B BROWN, PAUL FLICEK, YUTAO FU, DAMIAN KEEFE, EWAN BIRNEY, FRANCE DENOEUD, MARK GERSTEIN, ERIC D GREEN, PHILIPP KAPRANOV, ULAŞ KARAÖZ, RICHARD M MYERS, WILLIAM S NOBLE, ALEXANDRE REYMOND, JOEL ROZOWSKY, KEVIN STRUHL, ADAM SIEPEL, JOHN A STAMATOYANNOPOULOS, CHRISTOPHER M TAYLOR, JAMES TAYLOR, ROBERT E THURMAN, THOMAS D TULLIUS, STEFAN WASHIETL, DEYOU ZHENG, LAURA A LIEFER, KRIS A WETTERSTRAND, PETER J GOOD, ELISE A FEINGOLD, MARK S GUYER, FRANCIS S COLLINS, ELLIOTT H MARGULIES, GREGORY M COOPER, GEORGE ASIMENOS, DARYL J THOMAS, COLIN N DEWEY, ADAM SIEPEL, EWAN BIRNEY, DAMIAN KEEFE, MINMEI HOU, JAMES TAYLOR, SERGEY NIKOLAEV, JUAN I MONTOYA-BURGOS, ARI LÖYTYNOJA, SIMON WHELAN, FABIO PARDI, TIM MASSINGHAM, JAMES B BROWN, HAIYAN HUANG, NANCY R ZHANG, PETER BICKEL, IAN HOLMES, JAMES C MULLIKIN, ABEL URETA-VIDAL, BENEDICT PATEN, MICHAEL SERINGHAUS, DEANNA CHURCH, KATE ROSENBLOOM, W JAMES KENT, ERIC A STONE, NISC COMPARATIVE SEQUENCING PROGRAM, BAYLOR COLLEGE OF MEDICINE HUMAN GENOME SEQUENCING CENTER, WASHINGTON UNIVERSITY GENOME SEQUENCING CENTER, BROAD INSTITUTE, CHILDREN'S HOSPITAL OAKLAND RESEARCH INSTITUTE, MARK GERSTEIN, STYLIANOS E ANTONARAKIS, SERAFIM BATZOGLOU, NICK GOLDMAN, ROSS C HARDISON, DAVID HAUSSLER, WEBB MILLER, AND LIOR AND... PACHTER. **Identification and analysis of functional elements in 1human genome by the ENCODE pilot project**. *Nature*, **447**(7146):799–816, June 2007. 37

[104] C B LOZZIO AND B B LOZZIO. **Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome.** *Blood*, **45**(3):321–334, March 1975. 37

[105] A ARVEY, P AGIUS, W S NOBLE, AND C LESLIE. **Sequence and chromatin determinants of cell-type-specific transcription factor binding**. *Genome Research*, **22**(9):1723–1734, September 2012. 38

[106] A NATARAJAN, G G YARDIMCI, N C SHEFFIELD, G E CRAWFORD, AND U OHLER. **Predicting cell-type-specific gene expression from regions of open chromatin**. *Genome Research*, **22**(9):1711–1722, September 2012. 38

[107] M A Schaub, A P Boyle, A Kundaje, S Batzoglou, and M Snyder. **Linking disease associations with regulatory information in the human genome**. *Genome Research*, **22**(9):1748–1759, September 2012. 38, 95

[108] A P Boyle, E L Hong, M Hariharan, Y Cheng, M A Schaub, M Kasowski, K J Karczewski, J Park, B C Hitz, S Weng, J M Cherry, and M Snyder. **Annotation of functional variation in personal genomes using RegulomeDB**. *Genome Research*, **22**(9):1790–1797, September 2012. 38

[109] Bruce Beutler and Ernst Th Rietschel. **Innate immune sensing and its roots: the story of endotoxin.** *Nature reviews. Immunology*, **3**(2):169–176, February 2003. 39

[110] L Alexopoulou, A C Holt, R Medzhitov, and R A Flavell. **Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3.** *Nature*, **413**(6857):732–738, October 2001. 39

[111] Shizuo Akira, Satoshi Uematsu, and Osamu Takeuchi. **Pathogen Recognition and Innate Immunity**. *Cell*, **124**(4):783–801, February 2006. 39

[112] Daniel Ricklin, George Hajishengallis, Kun Yang, and John D Lambris. **Complement: a key system for immune surveillance and homeostasis**. *Nature Immunology*, **11**(9):785–797, August 2010. 39

[113] Jean-Laurent Casanova, Laurent Abel, and Lluis Quintana-Murci. **Human TLRs and IL-1Rs in Host Defense: Natural Insights from Evolutionary, Epidemiological, and Clinical Genetics**. *Annual review of immunology*, **29**(1):447–491, April 2011. 39

[114] Shizuo Akira and Kiyoshi Takeda. **Toll-like receptor signalling**. *Nature reviews. Immunology*, **4**(7):499–511, July 2004. 39, 42, 45

[115] Akiko Iwasaki and Ruslan Medzhitov. **Toll-like receptor control of the adaptive immune responses**. *Nature Immunology*, **5**(10):987–995, October 2004. 39, 42, 45

[116] Y. Asano and R.J. Hodes. **T cell regulation of B cell activation. T cells independently regulate the responses mediated by distinct B cell subpopulations.** *The Journal of experimental medicine*, **155**(5):1267–1276, 1982. 40

[117] J F MILLER and G F Mitchell. **Thymus and antigen-reactive cells.** *Transplantation reviews*, **1**:3–42, 1969.

[118] R L Coffman, B W Seymour, D A Lebman, D D Hiraki, J A Christiansen, B Shrader, H M Cherwinski, H F Savelkoul, F D Finkelman, and M W Bond. **The role of helper T cell products in mouse B cell differentiation and isotype regulation.** *Immunological Reviews*, **102**:5–28, February 1988. 40

[119] David M Mosser and Justin P Edwards. **Exploring the full spectrum of macrophage activation**. *Nature reviews. Immunology*, **8**(12):958–969, December 2008. 40

[120] T R Mosmann, H Cherwinski, M W Bond, M A Giedlin, and R L Coffman. **Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins.** *Journal of immunology (Baltimore, Md. : 1950)*, **136**(7):2348–2357, April 1986. 40, 44

[121] C A Thornton, J A Holloway, E J Popplewell, J K Shute, J Boughton, and J O Warner. **Fetal exposure to intact immunoglobulin E occurs via the gastrointestinal tract.** *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology*, **33**(3):306–311, March 2003. 40

[122] E A Nigro, A T Brini, E Soprana, A Ambrosi, D Dombrowicz, A G Siccardi, and L Vangelista. **Antitumor IgE Adjuvanticity: Key Role of Fc RI**. *The Journal of Immunology*, **183**(7):4530–4536, September 2009. 40

[123] H J Gould, G A Mackay, S N Karagiannis, C M O'Toole, P J Marsh, B E Daniel, L R Coney, V R Zurawski, M Joseph, M Capron, M Gilbert, G F Murphy, and R Korngold. **Comparison of IgE and IgG antibody-dependent cytotoxicity in vitro and in a SCID mouse xenograft model of ovarian carcinoma.** *European Journal of Immunology*, **29**(11):3527–3537, November 1999.

[124] Sophia N Karagiannis, Debra H Josephs, Panagiotis Karagiannis, Amy E Gilbert, Louise Saul, Sarah M Rudman, Tihomir Dodev, Alexander Koers, Philip J Blower, Christopher Corrigan, Andrew J Beavil, James F Spicer, Frank O Nestle, and Hannah J Gould. **Recombinant IgE antibodies for passive immunotherapy of solid tumours: from concept towards clinical application**. *Cancer Immunology, Immunotherapy*, **61**(9):1547–1564, December 2011. 40

[125] Ayelet Eppel, Michelle Cotterchio, and Steven Gallinger. **Allergies are associated with reduced pancreas cancer risk: A population-based case–control study in Ontario, Canada**. *International Journal of Cancer*, **121**(10):2241–2245, 2007. 40

[126] P K Mills, W L Beeson, G E Fraser, and R L Phillips. **Allergy and cancer: organ site-specific results from the Adventist Health Study.** *American journal of epidemiology*, **136**(3):287–295, August 1992. 40

[127] S C Liang, X Y Tan, D P Luxenberg, R Karim, K Dunussi-Joannopoulos, M Collins, and L A Fouser. **Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides**. *Journal of Experimental Medicine*, **203**(10):2271–2279, September 2006. 40

[128] Meredith M Curtis and Sing Sing Way. **Interleukin-17 in host defence against bacterial, mycobacterial and fungal pathogens**. *Immunology*, **126**(2):177–185, February 2009. 40

[129] Shimon Sakaguchi, Tomoyuki Yamaguchi, Takashi Nomura, and Masahiro Ono. **Regulatory T Cells and Immune Tolerance**. *Cell*, **133**(5):775–787, May 2008. 40

[130] Shane Crotty. **Follicular Helper CD4 T Cells (T FH)**. *Annual review of immunology*, **29**(1):621–663, April 2011. 40

[131] Shingo Nakayamada, Yuka Kanno, Hayato Takahashi, Dragana Jankovic, Kristina T Lu, Thomas A Johnson, Hong-wei Sun, Golnaz Vahedi, Ofir Hakim, Robin Handon, Pamela L Schwartzberg, Gordon L Hager, and John J O'Shea. **Early Th1 Cell Differentiation Is Marked by a Tfh Cell-like Transition**. *Immunity*, **35**(6):919–931, December 2011. 40, 45

[132] Kenneth J Oestreich, Sarah E Mohn, and Amy S Weinmann. **Molecular mechanisms that control the expression and activity of Bcl-6 in TH1 cells to regulate flexibility with a TFH-like gene profile**. *Nature Immunology*, **13**(4):405–411, March 2012. 40, 45

[133] Francesco Annunziato and Sergio Romagnani. **Heterogeneity of human effector CD4+ T cells.** *Arthritis research & therapy*, **11**(6):257, 2009. 41

[134] J J O'Shea and W E Paul. **Mechanisms Underlying Lineage Commitment and Plasticity of Helper CD4+ T Cells**. *Science*, **327**(5969):1098–1102, February 2010. 41

[135] J M Green, P J Noel, A I Sperling, T L Walunas, G S Gray, J A Bluestone, and C B Thompson. **Absence of B7-dependent responses in CD28-deficient mice.** *Immunity*, **1**(6):501–508, September 1994. 41

[136] Alison V Collins, Douglas W Brodie, Robert J C Gilbert, Andrea Iaboni, Raquel Manso-Sancho, Björn Walse, David I Stuart, P Anton van der Merwe, and Simon J Davis. **The interaction properties of costimulatory molecules revisited.** *Immunity*, **17**(2):201–210, August 2002. 41

[137] Lucy S K Walker and David M Sansom. **The emerging role of CTLA4 as a cell-extrinsic regulator of T cell responses**. *Nature reviews. Immunology*, **11**(12):852–863, November 2011. 41

[138] E A Tivol, F Borriello, A N Schweitzer, W P Lynch, J A Bluestone, and A H Sharpe. **Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4.** *Immunity*, **3**(5):541–547, November 1995. 42

[139] A N Schweitzer, F Borriello, R C Wong, A K Abbas, and A H Sharpe. **Role of costimulators in T cell differentiation: studies using antigen-presenting cells lacking expression of CD80 or CD86.** *Journal of immunology (Baltimore, Md. : 1950)*, **158**(6):2713–2722, March 1997. 42

[140] Toshinori Nakayama and Masakatsu Yamashita. **The TCR-mediated signaling pathways that control the direction of helper T cell differentiation**. *Seminars in Immunology*, **22**(5):303–309, October 2010. 42

[141] Alfonso Martín-Fontecha, Lindy L Thomsen, Sara Brett, Craig Gerard, Martin Lipp, Antonio Lanzavecchia, and Federica Sallusto. **Induced recruitment of NK cells to lymph nodes provides IFN-γ for TH1 priming**. *Nature Immunology*, **5**(12):1260–1265, November 2004. 42

[142] Maryam Afkarian, John R Sedy, Jianfei Yang, Nils G Jacobson, Nezih Cereb, Soo Y Yang, Theresa L Murphy, and Kenneth M Murphy. **T-bet is a STAT1-induced regulator of IL-12R expression in naïve CD4+ T cells.** *Nature Immunology*, **3**(6):549–557, June 2002. 42

[143] Edda G Schulz, Luca Mariani, Andreas Radbruch, and Thomas Höfer. **Sequential Polarization and Imprinting of Type 1 T Helper Lymphocytes by Interferon-γ and Interleukin-12**. *Immunity*, **30**(5):673–683, May 2009. 42

[144] C M Bacon, E F Petricoin, J R Ortaldo, R C Rees, A C Larner, J A Johnston, and J J O'Shea. **Interleukin 12 induces tyrosine phosphorylation and activation of STAT4 in human lymphocytes.** *Proceedings of the National Academy of Sciences of the United States of America*, **92**(16):7307–7311, August 1995. 42

[145] Jinfang Zhu, Dragana Jankovic, Andrew J Oler, Gang Wei, Suveena Sharma, Gangqing Hu, Liying Guo, Ryoji Yagi, Hidehiro Yamane, George Punkosdy, Lionel Feigenbaum, Keji Zhao, and William E Paul. **The Transcription Factor T-bet Is Induced by Multiple Pathways and Prevents an Endogenous Th2 Cell Program during Th1 Cell Responses**. *Immunity*, October 2012. 42

[146] K Schroder. **Interferon- : an overview of signals, mechanisms and functions**. *Journal of Leukocyte Biology*, **75**(2):163–189, December 2003. 42, 46, 47

[147] Kenji Tanigaki, Masayuki Tsuji, Norio Yamamoto, Hua Han, Jun Tsukada, Hiromasa Inoue, Masato Kubo, and Tasuku Honjo. **Regulation of alphabeta/gammadelta T cell lineage commitment and peripheral T cell responses by Notch/RBP-J signaling.** *Immunity*, **20**(5):611–622, May 2004. 43

[148] Derk Amsen, J Magarian Blander, Gap Ryol Lee, Kenji Tanigaki, Tasuku Honjo, and Richard A Flavell. **Instruction of distinct CD4 T helper cell fates by different notch ligands on antigen-presenting cells.** *Cell*, **117**(4):515–526, May 2004. 43, 44

[149] Lisa M Minter, Danielle M Turley, Pritam Das, Hyun Mu Shin, Ila Joshi, Rebecca G Lawlor, Ok Hyun Cho, Tanapat Palaga, Sridevi Gottipati, Janice C Telfer, Lisa Kostura, Abdul H Fauq, Katherine Simpson, Kimberly A Such, Lucio Miele, Todd E Golde, Stephen D Miller, and Barbara A Osborne. **Inhibitors of γ-secretase block in vivo and in vitro T helper type 1 polarization by preventing Notch upregulation of Tbx21**. *Nature Immunology*, **6**(7):680–688, May 2005. 43

[150] K Mark Ansel, Ivana Djuretic, Bogdan Tanasa, and Anjana Rao. **REGULATION OF TH2 DIFFERENTIATION AND Il4LOCUS ACCESSIBILITY**. *Annual review of immunology*, **24**(1):607–656, April 2006. 43, 53, 54

[151] K L Kemp, S D Levin, P J Bryce, and P L Stein. **Lck Mediates Th2 Differentiation through Effects on T-bet and GATA-3**. *The Journal of Immunology*, **184**(8):4178–4184, April 2010. 43

[152] C E Demeure, L P Yang, D G Byun, H Ishihara, N Vezzio, and G Delespesse. **Human naive CD4 T cells produce interleukin-4 at priming and acquire a Th2 phenotype upon repetitive stimulations in neutral conditions.** *European Journal of Immunology*, **25**(9):2722–2725, September 1995. 43

[153] M Moser and K M Murphy. **Dendritic cell regulation of TH1-TH2 development.** *Nature Immunology*, **1**(3):199–205, September 2000. 43

[154] Ahmed N Hegazy, Michael Peine, Caroline Helmstetter, Isabel Panse, Anja Fröhlich, Andreas Bergthaler, Lukas Flatz, Daniel D Pinschewer, Andreas Radbruch, and Max Löhning. **Interferons Direct Th2 Cell Reprogramming to Generate a Stable GATA-3+T-bet+ Cell Subset with Combined Th2 and Th1 Cell Functions**. *Immunity*, **32**(1):116–128, January 2010. 44

[155] Jinfang Zhu and William E Paul. **CD4+ T Cell Plasticity—Th2 Cells Join the Crowd**. *Immunity*, **32**(1):11–13, January 2010. 44

[156] Laurie E Harrington, Robin D Hatton, Paul R Mangan, Henrietta Turner, Theresa L Murphy, Kenneth M Murphy, and Casey T Weaver. **Interleukin 17–producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages**. *Nature Immunology*, **6**(11):1123–1132, October 2005. 44

[157] Yun Kyung Lee, Henrietta Turner, Craig L Maynard, James R Oliver, Dongquan Chen, Charles O Elson, and Casey T Weaver. **Late Developmental Plasticity in the T Helper 17 Lineage**. *Immunity*, **30**(1):92–107, January 2009. 44

[158] Richard G Jenner, Michael J Townsend, Ian Jackson, Kaiming Sun, Russell D Bouwman, Richard A Young, Laurie H Glimcher, and Graham M Lord. **The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes.** *Proceedings of the National Academy of Sciences of the United States of America*, **106**(42):17876–17881, October 2009. 45, 53, 57, 59, 79, 104, 110, 152

[159] Meghan A Koch, Glady's Tucker-Heard, Nikole R Perdue, Justin R Killebrew, Kevin B Urdahl, and Daniel J Campbell. **The transcription factor T-bet controls regulatory T cell homeostasis and function during type 1 inflammation**. *Nature Immunology*, **10**(6):595–602, May 2009. 45, 47, 58

[160] Claire L Langrish, Brent S McKenzie, Nicholas J Wilson, Rene de Waal Malefyt, Robert A Kastelein, and Daniel J Cua. **IL-12 and IL-23: master regulators of innate and adaptive immunity.** *Immunological Reviews*, **202**:96–105, December 2004. 45

[161] U Shu, M Kiniwa, C Y Wu, C Maliszewski, N Vezzio, J Hakimi, M Gately, and G Delespesse. **Activated T cells induce interleukin-12 production by monocytes via CD40-CD40 ligand interaction.** *European Journal of Immunology*, **25**(4):1125–1128, April 1995. 45

[162] U Grohmann, M L Belladonna, C Vacca, R Bianchi, F Fallarino, C Orabona, M C Fioretti, and P Puccetti. **Positive regulatory role of IL-12 in macrophages and modulation by IFN-gamma.** *Journal of immunology (Baltimore, Md. : 1950)*, **167**(1):221–227, July 2001. 45

[163] T Ohteki, T Fukao, K Suzue, C Maki, M Ito, M Nakamura, and S Koyasu. **Interleukin 12-dependent interferon gamma production by CD8alpha+ lymphoid dendritic cells.** *The Journal of experimental medicine*, **189**(12):1981–1986, June 1999. 45

[164] B Oppmann, R Lesley, B Blom, J C Timans, Y Xu, B Hunte, F Vega, N Yu, J Wang, K Singh, F Zonin, E Vaisberg, T Churakova, M Liu, D Gorman, J Wagner, S Zurawski, Y Liu, J S Abrams, K W Moore, D Rennick, R de Waal-Malefyt, C Hannum, J F Bazan, and R A Kastelein. **Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12.** *Immunity*, **13**(5):715–725, November 2000. 45

[165] Anand Balasubramani, Ryuta Mukasa, Robin D Hatton, and Casey T Weaver. **Regulation of the Ifng locus in the context of T-lineage specification and plasticity.** *Immunological Reviews*, **238**(1):216–232, November 2010. 45, 48

[166] Christi Parham, Madaline Chirica, Jacqueline Timans, Elena Vaisberg, Marilyn Travis, Jeanne Cheung, Stefan Pflanz, Rebecca Zhang, Komal P Singh, Felix Vega, Wayne To, Janet Wagner, Anne-Marie O'Farrell, Terrill McClanahan, Sandra Zurawski, Charles Hannum, Daniel Gorman, Donna M Rennick, Robert A Kastelein, Rene de Waal Malefyt, and Kevin W Moore. **A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R.** *Journal of immunology (Baltimore, Md. : 1950)*, **168**(11):5699–5708, June 2002. 45

[167] Wendy T Watford, Masato Moriguchi, Akio Morinobu, and John J O'Shea. **The biology of IL-12: coordinating innate and adaptive immune responses**. *Cytokine & Growth Factor Reviews*, **14**(5):361–368, October 2003. 45

[168] S E Ealick, W J Cook, S Vijay-Kumar, M Carson, T L Nagabhushan, P P Trotta, and C E Bugg. **Three-dimensional structure of recombinant human interferon-gamma.** *Science*, **252**(5006):698–702, May 1991. 46

[169] J Yang, T L Murphy, W Ouyang, and K M Murphy. **Induction of interferon-gamma production in Th1 CD4+ T cells: evidence for two distinct pathways for promoter activation.** *European Journal of Immunology*, **29**(2):548–555, February 1999. 46

[170] Anand Balasubramani, Yoichiro Shibata, Gregory E Crawford, Albert S Baldwin, Robin D Hatton, and Casey T Weaver. **Modular Utilization of Distal cis-Regulatory Elements Controls Ifng Gene Expression in T Cells Activated by Distinct Stimuli**. *Immunity*, **33**(1):35–47, July 2010. 46, 48, 49, 53, 59, 148, 186

[171] H Schindler, M B Lutz, M Röllinghoff, and C Bogdan. **The production of IFN-gamma by IL-12/IL-18-activated macrophages requires STAT4 signaling and is inhibited by IL-4.** *Journal of immunology (Baltimore, Md. : 1950)*, **166**(5):3075–3082, March 2001. 46

[172] P Bernabei, E M Coccia, L Rigamonti, M Bosticardo, G Forni, S Pestka, C D Krause, A Battistini, and F Novelli. **Interferon-gamma receptor 2 expression as the deciding factor in human T, B, and myeloid cell proliferation or death.** *Journal of Leukocyte Biology*, **70**(6):950–960, December 2001. 46

[173] M A Meraz, J M White, K C Sheehan, E A Bach, S J Rodig, A S Dighe, D H Kaplan, J K Riley, A C Greenlund, D Campbell, K Carver-Moore, R N DuBois, R Clark, M Aguet, and R D Schreiber. **Targeted disruption of the Stat1 gene in mice reveals unexpected physiologic specificity in the JAK-STAT signaling pathway.** *Cell*, **84**(3):431–442, February 1996. 46

[174] Xiaoyu Hu, Hao H Ho, Olivia Lou, Chisa Hidaka, and Lionel B Ivashkiv. **Homeostatic role of interferons conferred by inhibition of IL-1-mediated inflammation and tissue destruction.** *Journal of immunology (Baltimore, Md. : 1950)*, **175**(1):131–138, July 2005. 46

[175] Xiaoyu Hu and Lionel B Ivashkiv. **Cross-regulation of Signaling Pathways by Interferon-γ: Implications for Immune Responses and Autoimmune Diseases**. *Immunity*, **31**(4):539–550, October 2009. 46

[176] Z Wen, Z Zhong, and J E Darnell. **Maximal activation of transcription by Stat1 and Stat3 requires both tyrosine and serine phosphorylation.** *Cell*, **82**(2):241–250, July 1995. 47

[177] A J Loughlin, M N Woodroofe, and M L Cuzner. **Modulation of interferon-gamma-induced major histocompatibility complex class II and Fc receptor expression on isolated microglia by transforming growth factor-beta 1, interleukin-4, noradrenaline and glucocorticoids.** *Immunology*, **79**(1):125–130, May 1993. 47

[178] Xiaoyu Hu, Paul K Paik, Janice Chen, Anna Yarilina, Lisa Kockeritz, Theresa T Lu, James R Woodgett, and Lionel B Ivashkiv. **IFN-Îş Suppresses IL-10 Production and Synergizes with TLR2 by Regulating GSK3 and CREB/AP-1 Proteins**. *Immunity*, **24**(5):563–574, May 2006. 47

[179] Vanja Lazarevic, Xi Chen, Jae-Hyuck Shim, Eun-Sook Hwang, Eunjung Jang, Alexandra N Bolm, Mohamed Oukka, Vijay K Kuchroo, and Laurie H Glimcher. **T-bet represses TH17 differentiation by preventing Runx1-mediated activation of the gene encoding RORγt**. *Nature Immunology*, **12**(1):96–104, December 2010. 47, 57, 59

[180] Mohammed Soutto, Feng Zhang, Ben Enerson, Yingkai Tong, Mark Boothby, and Thomas M Aune. **A minimal IFN-gamma promoter confers Th1 selective expression.** *Journal of immunology (Baltimore, Md. : 1950)*, **169**(8):4205–4212, October 2002. 47, 142

[181] J Y Cho. **Identification of cooperative monomeric Brachyury sites conferring T-bet responsiveness to the proximal IFN- promoter**. *International Immunology*, **15**(10):1149–1160, October 2003. 47

[182] E L Pearce. **Control of Effector CD8+ T Cell Function by the Transcription Factor Eomesodermin**. *Science*, **302**(5647):1041–1043, November 2003. 47, 48, 58

[183] Mohammed Soutto, Weisong Zhou, and Thomas M Aune. **Cutting edge: distal regulatory elements are required to achieve selective expression of IFN-gamma in Th1/Tc1 effector cells.** *Journal of immunology (Baltimore, Md. : 1950)*, **169**(12):6664–6667, December 2002. 48

[184] Robin D Hatton, Laurie E Harrington, Rita J Luther, Therese Wakefield, Karen M Janowski, James R Oliver, Roger L Lallone, Kenneth M Murphy, and Casey T Weaver. **A Distal Conserved Sequence Element Controls Ifng Gene Expression by T Cells and NK Cells**. *Immunity*, **25**(5):717–729, November 2006. 48

[185] Masayuki Sekimata, Mercedes Pérez-Melgosa, Sara A Miller, Amy S Weinmann, Peter J Sabo, Richard Sandstrom, Michael O Dorschner, John A Stamatoyannopoulos, and Christopher B Wilson. **CCCTC-Binding Factor and the Transcription Factor T-bet Orchestrate T Helper 1 Cell-Specific Structure and Function at the Interferon-γ Locus**. *Immunity*, **31**(4):551–564, October 2009. 48

[186] Suzana Hadjur, Luke M Williams, Natalie K Ryan, Bradley S Cobb, Tom Sexton, Peter Fraser, Amanda G Fisher, and Matthias Merkenschlager. **Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus**. *Nature*, May 2009. 48

[187] M H Kaplan, Y L Sun, T Hoey, and M J Grusby. **Impaired IL-12 responses and enhanced development of Th2 cells in Stat4-deficient mice.** *Nature*, **382**(6587):174–177, July 1996. 48

[188] Susanne J Szabo, Brandon M Sullivan, Claudia Stemmann, Abhay R Satoskar, Barry P Sleckman, and Laurie H Glimcher. **Distinct effects of T-bet in TH1 lineage commitment and IFN-gamma production in CD4 and CD8 T cells.** *Science*, **295**(5553):338–342, January 2002. 48, 57, 58

[189] H Okamura, H Tsutsi, T Komatsu, M Yutsudo, A Hakura, T Tanimoto, K Torigoe, T Okura, Y Nukada, and K Hattori. **Cloning of a new cytokine that induces IFN-gamma production by T cells.** *Nature*, **378**(6552):88–91, November 1995. 48

[190] D Robinson, K Shibuya, A Mui, F Zonin, E Murphy, T Sana, S B Hartley, S Menon, R Kastelein, F Bazan, and A O'Garra. **IGIF does not drive Th1 development but synergizes with IL-12 for interferon-gamma production and activates IRAK and NFkappaB.** *Immunity*, **7**(4):571–581, October 1997. 49, 51, 155

[191] K Nakanishi, T Yoshimoto, H Tsutsui, and H Okamura. **Interleukin-18 regulates both Th1 and Th2 responses.** *Annual review of immunology*, **19**:423–474, 2001. 49

[192] A J Puren, G Fantuzzi, and C A Dinarello. **Gene expression, synthesis, and secretion of interleukin 18 and interleukin 1beta are differentially regulated in human blood mononuclear cells and mouse spleen cells.** *Proceedings of the National Academy of Sciences of the United States of America*, **96**(5):2256–2261, March 1999. 49

[193] Youichi Omoto, Kazuya Tokime, Keiichi Yamanaka, Koji Habe, Tatsuhiko Morioka, Ichiro Kurokawa, Hiroko Tsutsui, Kiyofumi Yamanishi, Kenji Nakanishi, and Hitoshi Mizutani. **Human mast cell chymase cleaves pro-IL-18 and generates a novel and biologically active IL-18 fragment.** *Journal of immunology (Baltimore, Md. : 1950)*, **177**(12):8315–8319, December 2006. 49

[194] William P Arend, Gaby Palmer, and Cem Gabay. **IL-1, IL-18, and IL-33 families of cytokines.** *Immunological Reviews*, **223**:20–38, June 2008. 49, 53

[195] D Xu, W L Chan, B P Leung, D Hunter, K Schulz, R W Carter, I B McInnes, J H Robinson, and F Y Liew. **Selective expression and functions of interleukin 18 receptor on T helper (Th) type 1 but not Th2 cells.** *The Journal of experimental medicine*, **188**(8):1485–1492, October 1998. 49, 50

[196] T L Born, E Thomassen, T A Bird, and J E Sims. **Cloning of a novel receptor subunit, AcPL, required for interleukin-18 signaling.** *The Journal of biological chemistry*, **273**(45):29445–29450, November 1998. 49

[197] Heidi Cheung, Nien-Jung Chen, Zhaodan Cao, Nobuyuki Ono, Pamela S Ohashi, and Wen-Chen Yeh. **Accessory protein-like is essential for IL-18-mediated signaling.** *Journal of immunology (Baltimore, Md. : 1950)*, **174**(9):5351–5357, May 2005. 49, 50

[198] D Wald, M Commane, G R Stark, and X Li. **IRAK and TAK1 are required for IL-18-mediated signaling.** *European Journal of Immunology*, **31**(12):3747–3754, December 2001. 49

[199] J Yang, H Zhu, T L Murphy, W Ouyang, and K M Murphy. **IL-18-stimulated GADD45 beta required in cytokine-induced, but not TCR-induced, IFN-gamma production.** *Nature Immunology*, **2**(2):157–164, February 2001. 49, 149

[200] Julie Chaix, Marlowe S Tessmer, Kasper Hoebe, Nicolas Fuséri, Bernhard Ryffel, Marc Dalod, Lena Alexopoulou, Bruce Beutler, Laurent Brossay, Eric Vivier, and Thierry Walzer. **Cutting edge: Priming of NK cells by IL-18.** *The Journal of Immunology*, **181**(3):1627–1631, August 2008. 50

[201] K Takeda, H Tsutsui, T Yoshimoto, O Adachi, N Yoshida, T Kishimoto, H Okamura, K Nakanishi, and S Akira. **Defective NK cell activity and Th1 response in IL-18-deficient mice.** *Immunity*, **8**(3):383–390, March 1998. 50

[202] K Hoshino, H Tsutsui, T Kawai, K Takeda, K Nakanishi, Y Takeda, and S Akira. **Cutting edge: generation of IL-18 receptor-deficient mice: evidence for IL-1 receptor-related protein as an essential IL-18 binding receptor.** *Journal of immunology (Baltimore, Md. : 1950)*, **162**(9):5041–5044, May 1999. 50

[203] T Yoshimoto, H Mizutani, H Tsutsui, N Noben-Trauth, K Yamanaka, M Tanaka, S Izumi, H Okamura, W E Paul, and K Nakanishi. **IL-18 induction of IgE: dependence on CD4+ T cells, IL-4 and STAT6.** *Nature Immunology*, **1**(2):132–137, August 2000. 50

[204] Qing Yu, Hua-Chen Chang, Ayele-Nati N Ahyi, and Mark H Kaplan. **Transcription factor-dependent chromatin remodeling of Il18r1 during Th1 and Th2 differentiation.** *The Journal of Immunology*, **181**(5):3346–3352, September 2008. 51, 52, 148, 152, 155, 202

[205] T Yoshimoto, K Takeda, T Tanaka, K Ohkusu, S Kashiwamura, H Okamura, S Akira, and K Nakanishi. **IL-12 upregulates IL-18 receptor expression on T cells, Th1 cells, and B cells: synergism with IL-18 for IFN-gamma production.** *Journal of immunology (Baltimore, Md. : 1950)*, **161**(7):3400–3407, October 1998.

[206] R B Smeltz, J Chen, J Hu-Li, and E M Shevach. **Regulation of interleukin (IL)-18 receptor alpha chain expression on CD4(+) T cells during T helper (Th)1/Th2 differentiation. Critical downregulatory role of IL-4.** *The Journal of experimental medicine*, **194**(2):143–153, July 2001. 51, 149, 155

[207] H Hata. **IL-18 together with anti-CD3 antibody induces human Th1 cells to produce Th1- and Th2-cytokines and IL-8**. *International Immunology*, **16**(12):1733–1739, October 2004. 51

[208] T Sugimoto. **Interleukin 18 Acts on Memory T Helper Cells Type 1 to Induce Airway Inflammation and Hyperresponsiveness in a Naive Host Mouse**. *Journal of Experimental Medicine*, **199**(4):535–545, February 2004. 51

[209] S J Lalor, L S Dungan, C E Sutton, S A Basdeo, J M Fletcher, and K H G Mills. **Caspase-1-Processed Cytokines IL-1 and IL-18 Promote IL-17 Production by and CD4 T Cells That Mediate Autoimmunity**. *The Journal of Immunology*, **186**(10):5738–5748, May 2011. 51

[210] Ilona Gutcher, Eduard Urich, Karina Wolter, Marco Prinz, and Burkhard Becher. **Interleukin 18–independent engagement of interleukin 18 receptor-*alpha* is required for autoimmune inflammation**. *Nature Immunology*, **7**(9):946–953, August 2006. 52

[211] G Pan, P Risser, W Mao, D T Baldwin, A W Zhong, E Filvaroff, D Yansura, L Lewis, C Eigenbrot, W J Henzel, and R Vandlen. **IL-1H, an interleukin 1-related protein that binds IL-18 receptor/IL-1Rrp.** *Cytokine*, **13**(1):1–7, January 2001. 52

[212] M Nakahira, M Tomura, M Iwasaki, H J Ahn, Y Bian, T Hamaoka, T Ohta, M Kurimoto, and H Fujiwara. **An absolute requirement for STAT4 and a role for IFN-gamma as an amplifying factor in IL-12 induction of the functional IL-18 receptor complex.** *Journal of immunology (Baltimore, Md. : 1950)*, **167**(3):1306–1312, August 2001. 52

[213] Vivian T Thieu, Qing Yu, Hua-Chen Chang, Norman Yeh, Evelyn T Nguyen, Sarita Sehra, and Mark H Kaplan. **Signal Transducer and Activator of Transcription 4 Is Required for the Transcription Factor T-bet to Promote T Helper 1 Cell-Fate Determination**. *Immunity*, **29**(5):679–690, November 2008. 53, 59, 148, 149, 202

[214] Karen A Hunt, Alexandra Zhernakova, Graham Turner, Graham A R Heap, Lude Franke, Marcel Bruinenberg, Jihane Romanos, Lotte C Dinesen, Anthony W Ryan, Davinder Panesar, Rhian Gwilliam, Fumihiko Takeuchi, William M McLaren, Geoffrey K T Holmes, Peter D Howdle, Julian R F Walters, David S Sanders, Raymond J Playford, Gosia Trynka, Chris J J Mulder, M Luisa Mearin, Wieke H M Verbeek, Valerie Trimble, Fiona M Stevens, Colm O'Morain, Nicholas P Kennedy, Dermot Kelleher, Daniel J Pennington, David P Strachan, Wendy L McArdle, Charles A Mein, Martin C Wapenaar, Panos Deloukas, Ralph McGinnis, Ross McManus, Cisca Wijmenga, and David A van Heel. **Newly identified genetic risk variants for celiac disease related to the immune response**. *Nature Genetics*, **40**(4):395–402, March 2008. 53, 137, 204

[215] W Zheng and R A Flavell. **The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells.** *Cell*, **89**(4):587–596, May 1997. 54, 60

[216] Jinfang Zhu, Booki Min, Jane Hu-Li, Cynthia J Watson, Alex Grinberg, Qi Wang, Nigel Killeen, Joseph F Urban, Liying Guo, and William E Paul. **Conditional deletion of Gata3 shows its essential function in TH1-TH2 responses**. *Nature Immunology*, **5**(11):1157–1165, October 2004. 54

[217] S Agarwal, O Avni, and A Rao. **Cell-type-restricted binding of the transcription factor NFAT to a distal IL-4 enhancer in vivo.** *Immunity*, **12**(6):643–652, June 2000. 54

[218] M D Siegel, D H Zhang, P Ray, and A Ray. **Activation of the interleukin-5 promoter by cAMP in murine EL-4 cells requires the GATA-3 and CLE0 elements.** *The Journal of biological chemistry*, **270**(41):24548–24555, October 1995. 54

[219] H Kishikawa, J Sun, A Choi, S C Miaw, and I C Ho. **The cell type-specific expression of the murine IL-13 gene is regulated by GATA-3.** *Journal of immunology (Baltimore, Md. : 1950)*, **167**(8):4414–4420, October 2001. 54

[220] L O Tykocinski. **A Critical Control Element for Interleukin-4 Memory Expression in T Helper Lymphocytes**. *Journal of Biological Chemistry*, **280**(31):28177–28185, June 2005. 54

[221] Anne S Hutchins, Alan C Mullen, Hubert W Lee, Kara J Sykes, Frances A High, Brian D Hendrich, Adrian P Bird, and Steven L Reiner. **Gene silencing quantitatively controls the function of a developmental trans-activator.** *Molecular cell*, **10**(1):81–91, July 2002. 54

[222] Gap Ryol Lee, Patrick E Fields, Thomas J Griffin, and Richard A Flavell. **Regulation of the Th2 cytokine locus by a locus control region.** *Immunity*, **19**(1):145–153, July 2003. 54

[223] Charalampos G Spilianakis and Richard A Flavell. **Long-range intrachromosomal interactions in the T helper type 2 cytokine locus**. *Nature Immunology*, **5**(10):1017–1027, September 2004. 55

[224] K Nelms, A D Keegan, J Zamorano, J J Ryan, and W E Paul. **The IL-4 receptor: signaling mechanisms and biologic functions.** *Annual review of immunology*, **17**:701–738, 1999. 55

[225] N I Obiri, W Debinski, W J Leonard, and R K Puri. **Receptor for interleukin 13. Interaction with interleukin 4 by a mechanism that does not involve the common gamma chain shared by receptors for interleukins 2, 4, 7, 9, and 15.** *The Journal of biological chemistry*, **270**(15):8797–8804, April 1995. 55

[226] G J McKenzie, A Bancroft, R K Grencis, and A N McKenzie. **A distinct role for interleukin-13 in Th2-cell-mediated immune responses.** *Current biology : CB*, **8**(6):339–342, March 1998. 55

[227] M G Chiaramonte, D D Donaldson, A W Cheever, and T A Wynn. **An IL-13 inhibitor blocks the development of hepatic fibrosis during a T-helper type 2-dominated inflammatory response.** *The Journal of clinical investigation*, **104**(6):777–785, September 1999. 55

[228] M Kopf, F Brombacher, P D Hodgkin, A J Ramsay, E A Milbourne, W J Dai, K S Ovington, C A Behm, G Köhler, I G Young, and K I Matthaei. **IL-5-deficient mice have a developmental defect in CD5+ B-1 cells and lack eosinophilia but have normal antibody and cytotoxic T cell responses.** *Immunity*, **4**(1):15–24, January 1996. 56

[229] S J Szabo, S T Kim, G L Costa, X Zhang, C G Fathman, and L H Glimcher. **A novel transcription factor, T-bet, directs Th1 lineage commitment.** *Cell*, **100**(6):655–669, March 2000. 56, 57, 120

[230] Kamel El Omari, Julie De Mesmaeker, Dimple Karia, Helen Ginn, Shoumo Bhattacharya, and Erika J Mancini. **Structure of the DNA-bound T-box domain of human TBX1, a transcription factor associated with the DiGeorge syndrome**. *Proteins: Structure, Function, and Bioinformatics*, **80**(2):655–660, November 2011. 56

[231] Miquel Coll, Jonathan G Seidman, and Christoph W Müller. **Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome.** *Structure (London, England : 1993)*, **10**(3):343–356, March 2002. 56

[232] Christian U Stirnimann, Denis Ptchelkine, Clemens Grimm, and Christoph W Müller. **Structural Basis of TBX5–DNA Recognition: The T-Box Domain in Its DNA-Bound and -Unbound Form**. *Journal of Molecular Biology*, **400**(1):71–81, July 2010. 56

[233] C W Müller and B G Herrmann. **Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor.** *Nature*, **389**(6653):884–888, October 1997. 56

[234] Aditi Kanhere, Arnulf Hertweck, Urvashi Bhatia, M Refik Gökmen, Esperanza Perucha, Ian Jackson, Graham M Lord, and Richard G Jenner. **T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements**. *Nature Communications*, **3**:1268, December 2012. 56, 57, 59, 69, 71, 95, 96, 99, 103, 104, 121, 188, 193, 198

[235] Mitsuteru Akahoshi, Kazuhiko Obara, Tomomitsu Hirota, Akira Matsuda, Koichi Hasegawa, Naomi Takahashi, Makiko Shimizu, Kazuko Nakashima, Lei Cheng, Satoru Doi, Hiroshi Fujiwara, Akihiko Miyatake, Kimie Fujita, Noritaka Higashi, Masami Taniguchi, Tadao Enomoto, Xiao-Quan Mao, Hitoshi Nakashima, Chaker N Adra, Yusuke Nakamura, Mayumi Tamari, and Taro Shirakawa. **Functional promoter polymorphism in the TBX21 gene associated with aspirin-induced asthma**. *Human Genetics*, **117**(1):16–26, April 2005. 57

[236] Yuka Sasaki, Kenji Ihara, Nobuo Matsuura, Hitoshi Kohno, Seiho Nagafuchi, Ryuichi Kuromaru, Koichi Kusuhara, Ryu Takeya, Timothy Hoey, Hideki Sumimoto, and Toshiro Hara. **Identification of a novel type 1 diabetes susceptibility gene, T-bet**. *Human Genetics*, **115**(3), July 2004. 57

[237] G M Lord, R M Rao, H Choe, B M Sullivan, A H Lichtman, F W Luscinskas, and L H Glimcher. **T-bet is required for optimal proinflammatory CD4+ T-cell trafficking**. *Blood*, **106**(10):3432–3439, November 2005. 57

[238] Kristin M Beima, Michael M Miazgowicz, Megan D Lewis, Pearlly S Yan, Tim H-M Huang, and Amy S Weinmann. **T-bet binding to newly identified target gene promoters is cell type-independent but results in variable context-dependent functional effects.** *The Journal of biological chemistry*, **281**(17):11992–12000, April 2006. 57

[239] Michael J Townsend, Amy S Weinmann, Jennifer L Matsuda, Rachelle Salomon, Peggy J Farnham, Christine A Biron, Laurent Gapin, and Laurie H Glimcher. **T-bet regulates the terminal maturation and homeostasis of NK and Valpha14i NKT cells.** *Immunity*, **20**(4):477–494, April 2004. 57

[240] Brandon M Sullivan, Ousman Jobe, Vanja Lazarevic, Kristine Vasquez, Roderick Bronson, Laurie H Glimcher, and Igor Kramnik. **Increased susceptibility of mice lacking T-bet to infection with Mycobacterium tuberculosis correlates with increased IL-10 and decreased IFN-gamma production.** *Journal of immunology (Baltimore, Md. : 1950)*, **175**(7):4593–4602, October 2005. 57

[241] J H Esensten, M R Lee, L H Glimcher, and J A Bluestone. **T-bet-Deficient NOD Mice Are Protected from Diabetes Due to Defects in Both T Cell and Innate Immune System Function**. *The Journal of Immunology*, **183**(1):75–82, June 2009. 57

[242] Stanford L Peng, Susanne J Szabo, and Laurie H Glimcher. **T-bet regulates IgG class switching and pathogenic autoantibody production.** *Proceedings of the National Academy of Sciences of the United States of America*, **99**(8):5545–5550, April 2002. 58

[243] J Wang. **Transcription factor T-bet regulates inflammatory arthritis through its function in dendritic cells**. *The Journal of clinical investigation*, **116**(2):414–421, January 2006. 58

[244] Wendy S Garrett, Graham M Lord, Shivesh Punit, Geancarlo Lugo-Villarino, Sarkis K Mazmanian, Susumu Ito, Jonathan N Glickman, and Laurie H Glimcher. **Communicable Ulcerative Colitis Induced by T-bet Deficiency in the Innate Immune System**. *Cell*, **131**(1):33–45, October 2007. 58, 64

[245] Nick Powell, Alan W Walker, Emilie Stolarczyk, James B Canavan, M Refik Gökmen, Ellen Marks, Ian Jackson, Ahmed Hashim, Michael A Curtis, Richard G Jenner, Jane K Howard, Julian Parkhill, Thomas T MacDonald, and Graham M Lord. **The Transcription Factor T-bet Regulates Intestinal Inflammation Mediated by Interleukin-7 Receptor+ Innate Lymphoid Cells**. *Immunity*, **37**(4):674–684, October 2012. 58, 64, 80

[246] Brandon M Sullivan, Amy Juedes, Susanne J Szabo, Matthias von Herrath, and Laurie H Glimcher. **Antigen-driven effector CD8 T cell function regulated by T-bet.** *Proceedings of the National Academy of Sciences of the United States of America*, **100**(26):15818–15823, December 2003. 58

[247] S A Miller, A C Huang, M M Miazgowicz, M M Brassil, and A S Weinmann. **Coordinated but physically separable interaction with H3K27-demethylase and H3K4-methyltransferase activities are required for T-box protein-mediated activation of developmental gene expression**. *Genes & Development*, **22**(21):2980–2993, November 2008. 58

[248] M D Lewis, S A Miller, M M Miazgowicz, K M Beima, and A S Weinmann. **T-bet's Ability To Regulate Individual Target Genes Requires the Conserved T-Box Domain To Recruit Histone Methyltransferase Activity and a Separate Family Member-Specific Transactivation Domain**. *Molecular and Cellular Biology*, **27**(24):8510–8521, November 2007. 58

[249] E S Hwang. **T Helper Cell Fate Specified by Kinase-Mediated Interaction of T-bet with GATA-3**. *Science*, **307**(5708):430–433, January 2005. 59

[250] E S Hwang. **IL-2 production in developing Th1 cells is regulated by heterodimerization of RelA and T-bet and requires T-bet serine residue 508**. *Journal of Experimental Medicine*, **202**(9):1289–1300, October 2005. 59

[251] Ivana M Djuretic, Ditsa Levanon, Varda Negreanu, Yoram Groner, Anjana Rao, and K Mark Ansel. **Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells**. *Nature Immunology*, **8**(2):145–153, December 2006. 59

[252] P P Pandolfi, M E Roth, A Karis, M W Leonard, E Dzierzak, F G Grosveld, J D Engel, and M H Lindenbaum. **Targeted disruption of the GATA3 gene causes severe abnormalities in the nervous system and in fetal liver haematopoiesis.** *Nature Genetics*, **11**(1):40–44, September 1995. 59

[253] C N Ting, M C Olson, K P Barton, and J M Leiden. **Transcription factor GATA-3 is required for development of the T-cell lineage.** *Nature*, **384**(6608):474–478, December 1996. 59, 60

[254] J A Lowry and W R Atchley. **Molecular evolution of the GATA family of transcription factors: conservation within the DNA-binding domain.** *Journal of molecular evolution*, **50**(2):103–115, February 2000. 59

[255] A Newton. **The N-terminal Zinc Finger of the Erythroid Transcription Factor GATA-1 Binds GATC Motifs in DNA**. *Journal of Biological Chemistry*, **276**(38):35794–35801, July 2001. 59

[256] Darren L Bates, Yongheng Chen, Grace Kim, Liang Guo, and Lin Chen. **Crystal Structures of Multiple GATA Zinc Fingers Bound to DNA Reveal New Insights into DNA Recognition and Self-Association by GATA**. *Journal of Molecular Biology*, **381**(5):1292–1306, September 2008. 60

[257] A Y Barakat, J B D'Albora, M M Martin, and P A Jose. **Familial nephrosis, nerve deafness, and hypoparathyroidism.** *The Journal of pediatrics*, **91**(1):61–64, July 1977. 60

[258] M Yamamoto, L J Ko, M W Leonard, H Beug, S H Orkin, and J D Engel. **Activity and tissue-specific expression of the transcription factor NF-E1 multigene family.** *Genes & Development*, **4**(10):1650–1662, October 1990. 60

[259] L J Ko, M Yamamoto, M W Leonard, K M George, P Ting, and J D Engel. **Murine and human T-lymphocyte GATA-3 factors mediate transcription through a cis-regulatory element within the human T-cell receptor delta gene enhancer.** *Molecular and Cellular Biology*, **11**(5):2778–2784, May 1991. 60

[260] W Ouyang, M Löhning, Z Gao, M Assenmacher, S Ranganath, A Radbruch, and K M Murphy. **Stat6-independent GATA-3 autoactivation directs IL-4-independent Th2 development and commitment.** *Immunity*, **12**(1):27–37, January 2000. 61

[261] K Kohu, H Ohmori, W F Wong, D Onda, T Wakoh, S Kon, M Yamashita, T Nakayama, M Kubo, and M Satake. **The Runx3 Transcription Factor Augments Th1 and Down-Modulates Th2 Phenotypes by Interacting with and Attenuating GATA3**. *The Journal of Immunology*, **183**(12):7817–7824, December 2009. 61

[262] Pierre-Yves Mantel, Harmjan Kuipers, Onur Boyman, Claudio Rhyner, Nadia Ouaked, Beate Rückert, Christian Karagiannidis, Bart N Lambrecht, Rudolf W Hendriks, Reto Crameri, Cezmi A Akdis, Kurt Blaser, and Carsten B Schmidt-Weber. **GATA3-Driven Th2 Responses Inhibit TGF-Î§1–Induced FOXP3 Expression and the Formation of Regulatory T Cells**. *PLoS Biology*, **5**(12):e329, 2007. 61

[263] Y Nakamura, O Ghaffar, R Olivenstein, R A Taha, A Soussi-Gounni, D H Zhang, A Ray, and Q Hamid. **Gene expression of the GATA-3 transcription factor is increased in atopic asthma.** *J Allergy Clin Immunol*, **103**(2 Pt 1):215–222, February 1999. 61

[264] Yutaka Nakamura, Pota Christodoulopoulos, Lisa Cameron, Erin Wright, Francois Lavigne, Masao Toda, Shigeo Muro, Anuradha Ray, David H Eidelman, Eleanor Minshall, and Qutayba Hamid. **Upregulation of the transcription factor GATA-3 in upper airway mucosa after in vivo and in vitro allergen challenge**. *Journal of Allergy and Clinical Immunology*, **105**(6):1146–1152, June 2000. 61

[265] W Yan, Q J Cao, R B Arenas, B Bentley, and R Shao. **GATA3 Inhibits Breast Cancer Metastasis through the Reversal of Epithelial-Mesenchymal Transition**. *Journal of Biological Chemistry*, **285**(18):14042–14051, April 2010. 61

[266] Nam K Yoon, Erin L Maresh, Dejun Shen, Yahya Elshimali, Sophia Apple, Steve Horvath, Vei Mah, Shikha Bose, David Chia, Helena R Chang, and Lee Goodglick. **Higher levels of GATA3 predict better survival in women with breast cancer**. *Human Pathology*, **41**(12):1794–1801, December 2010. 61

[267] AM James Shapiro, Jonathan RT Lakey, Edmond A Ryan, Gregory S Korbutt, Ellen Toth, Garth L Warnock, Norman M Kneteman, and Ray V Rajotte. **Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen**. *The New England journal of medicine*, **343**(4):230–238, 2000. 62

[268] John A Todd. **Etiology of Type 1 Diabetes**. *Immunity*, **32**(4):457–467, April 2010. 62, 63

[269] Sergey Nejentsev, Joanna M M Howson, Neil M Walker, Jeffrey Szeszko, Sarah F Field, Helen E Stevens, Pamela Reynolds, Matthew Hardy, Erna King, Jennifer Masters, John Hulme, Lisa M Maier, Deborah Smyth, Rebecca Bailey, Jason D Cooper, Gloria Ribas, R Duncan Campbell, David G Clayton, John A Todd, Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, John A Todd, Peter Donnelly, Jeffrey C Barrett, Paul R Burton, Dan Davison, Peter Donnelly, Doug Easton, David Evans, Hin-Tak Leung, Jonathan L Marchini, Andrew P Morris, Chris C A Spencer, Martin D Tobin, Lon R Cardon, David G Clayton, Antony P Attwood, James P Boorman, Barbara Cant, Ursula Everson, Judith M Hussey, Jennifer D Jolley, Alexandra S Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V Prowse, Helen E Stevens, Niall C Taylor, Graham R Walters, Neil M Walker, Nicholas A Watkins, Thilo Winzer, John A Todd, Willem H Ouwehand, Richard W Jones, Wendy L McArdle, Susan M Ring, David P Strachan, Marcus Pembrey, Gerome Breen, David St Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K Green, Detelina Grozeva, Marian L Hamshere, Peter A Holmans, Ian R Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C O'Donovan, Michael J Owen, Nick Craddock, David A Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H Young, I Nicol Ferrier, Stephen G Ball, Anthony J Balmforth, Jennifer H Barrett, D Timothy Bishop, Mark M Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S Hall, Peter S Braund, Paul R Burton, Richard J Dixon, Massimo Mangino, Suzanne Stevens, Martin D Tobin, John R Thompson, Nilesh J Samani, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W Lees, Elaine R Nimmo, Jack Satsangi, Sheila A Fisher, Alastair Forbes, Cathryn M Lewis, Clive M Onnie, Natalie J Prescott, Jeremy Sanderson, Christopher G Mathew, Jamie Barbour, M Khalid Mohiuddin, Catherine E Todhunter, John C Mansfield, Tariq Ahmad, Fraser R Cummings, Derek P Jewell, John Webster, Morris J Brown, David G Clayton, G Mark Lathrop, John Connell, Anna Dominiczak, Nilesh J Samani, Carolina A Braga, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L Lee, Patricia B Munroe, Stephen J Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, Ian N Bruce, Hannah Donovan, Steve Eyre, Paul D Gilbert, Samantha L Hider, Anne M Hinks, Sally L John, Catherine Potter, Alan J Silman, Deborah P M Symmons, Wendy Thomson, Jane Worthington, David G Clayton, David B Dunger, Sarah Nutland, Helen E Stevens, Neil M Walker, Barry Widmer, John A Todd, Timothy M Frayling, Rachel M Freathy, Hana Lango, John R B Perry, Beverley M Shields, Michael N Weedon, Andrew T Hattersley, Graham A Hitman, Mark Walker, Kate S Elliott, Christopher J Groves, Cecilia M Lindgren, Nigel W Rayner, Nicholas J Timpson, Eleftheria Zeggini, Mark I McCarthy, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V S Hill, Linda A Bradbury, Claire Farrar, Jennifer J Pointon, Paul Wordsworth, Matthew A Brown, Jayne A Franklyn, Joanne M Heward, Matthew J Simmonds, Stephen C L Gough, Sheila Seal, Michael R Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A Rockett, Dominic P Kwiatkowski, Claire Bryan, Suzannah J Bumpstead, Amy Chaney, Kate Downes, Jilur Ghori, Rhian Gwilliam, Sarah E Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, David Withers, Panos Deloukas, Hin-Tak Leung, Sarah Nutland, Helen E Stevens, Neil M Walker, John A Todd, and ... Easton. **Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A**. *Nature*, **450**(7171):887–892, November 2007. 62

[270] Christopher E Lowe, Jason D Cooper, Todd Brusko, Neil M Walker, Deborah J Smyth, Rebecca Bailey, Kirsi Bourget, Vincent Plagnol, Sarah Field, Mark Atkinson, David G Clayton, Linda S Wicker, and John A Todd. **Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes**. *Nature Genetics*, **39**(9):1074–1082, August 2007. 62

[271] Jeffrey C Barrett, David G Clayton, Patrick Concannon, Beena Akolkar, Jason D Cooper, Henry A Erlich, Cécile Julier, Grant Morahan, Jørn Nerup, Concepcion Nierras, Vincent Plagnol, Flemming Pociot, Helen Schuilenburg, Deborah J Smyth, Helen Stevens, John A Todd, Neil M Walker, and Stephen S Rich. **Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes**. *Nature Genetics*, **41**(6):703–707, May 2009. 62

[272] Sergey Nejentsev, Neil Walker, David Riches, Michael Egholm, and John A Todd. **Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science*, **324**(5925):387–389, April 2009. 62

[273] Deborah J Smyth, Vincent Plagnol, Neil M Walker, Jason D Cooper, Kate Downes, Jennie H M Yang, Joanna M M Howson, Helen Stevens, Ross McManus, Cisca Wijmenga, Graham A Heap, Patrick C Dubois, David G Clayton, Karen A Hunt, David A van Heel, and John A Todd. **Shared and distinct genetic variants in type 1 diabetes and celiac disease.** *The New England journal of medicine*, **359**(26):2767–2777, December 2008. 62, 65

[274] Kendra Vehik and Dana Dabelea. **The changing epidemiology of type 1 diabetes: why is it going through the roof?** *Diabetes/Metabolism Research and Reviews*, **27**(1):3–13, October 2010. 63

[275] Jeffrey C Barrett, James C Lee, Charles W Lees, Natalie J Prescott, Carl A Anderson, Anne Phillips, Emma Wesley, Kirstie Parnell, Hu Zhang, Hazel Drummond, Elaine R Nimmo, Dunecan Massey, Kasia Blaszczyk, Timothy Elliott, Lynn Cotterill, Helen Dallal, Alan J Lobo, Craig Mowat, Jeremy D Sanderson, Derek P Jewell, William G Newman, Cathryn Edwards, Tariq Ahmad, John C Mansfield, Jack Satsangi, Miles Parkes, Christopher G Mathew,

Peter Donnelly, Leena Peltonen, Jenefer M Blackwell, Elvira Bramon, Matthew A Brown, Juan P Casas, Aiden Corvin, Nicholas Craddock, Panos Deloukas, Audrey Duncanson, Janusz Jankowski, Hugh S Markus, Christopher G Mathew, Mark I McCarthy, Colin N A Palmer, Robert Plomin, Anna Rautanen, Stephen J Sawcer, Nilesh Samani, Richard C Trembath, Ananth C Viswanathan, Nicholas Wood, Chris C A Spencer, Jeffrey C Barrett, Céline Bellenguez, Daniel Davison, Colin Freeman, Amy Strange, Peter Donnelly, Cordelia Langford, Sarah E Hunt, Sarah Edkins, Rhian Gwilliam, Hannah Blackburn, Suzannah J Bumpstead, Serge Dronov, Matthew Gillman, Emma Gray, Naomi Hammond, Alagurevathi Jayakumar, Owen T McCann, Jennifer Liddle, Marc L Perez, Simon C Potter, Radhi Ravindrarajah, Michelle Ricketts, Matthew Waller, Paul Weston, Sara Widaa, Pamela Whittaker, Panos Deloukas, Leena Peltonen, Christopher G Mathew, Jenefer M Blackwell, Matthew A Brown, Aiden Corvin, Mark I McCarthy, Chris C A Spencer, Antony P Attwood, Jonathan Stephens, Jennifer Sambrook, Willem H Ouwehand, Wendy L McArdle, Susan M Ring, and David P Strachan. **Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region**. *Nature Genetics*, **41**(12):1330–1334, November 2009. 63

[276] J P Hugot, M Chamaillard, H Zouali, S Lesage, J P Cézard, J Belaiche, S Almer, C Tysk, C A O'Morain, M Gassull, V Binder, Y Finkel, A Cortot, R Modigliani, P Laurent-Puig, C Gower-Rousseau, J Macry, J F Colombel, M Sahbatou, and G Thomas. **Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease.** *Nature*, **411**(6837):599–603, May 2001. 63

[277] Y Ogura, D K Bonen, N Inohara, D L Nicolae, F F Chen, R Ramos, H Britton, T Moran, R Karaliuskas, R H Duerr, J P Achkar, S R Brant, T M Bayless, B S Kirschner, S B Hanauer, G Nuñez, and J H Cho. **A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease.** *Nature*, **411**(6837):603–606, May 2001. 63

[278] C W Lees, J C Barrett, M Parkes, and J Satsangi. **New IBD genetics: common pathways with other diseases**. *Gut*, **60**(12):1739–1753, November 2011. 63

[279] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, Jonah Essers, Mitja Mitrovic, Kaida Ning, Isabelle Cleynen, Emilie Theatre, Sarah L Spain, Soumya Raychaudhuri, Philippe Goyette, Zhi Wei, Clara Abraham, Jean-Paul Achkar, Tariq Ahmad, Leila Amininejad, Ashwin N Ananthakrishnan, Vibeke Andersen, Jane M Andrews, Leonard Baidoo, Tobias Balschun, Peter A Bampton, Alain Bitton, Gabrielle Boucher, Stephan Brand, Carsten Büning, Ariella Cohain, Sven Cichon, Mauro D'Amato, Dirk De Jong, Kathy L Devaney, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Lynnette R Ferguson, Denis Franchimont, Karin Fransen, Richard Gearry, Michel Georges, Christian Gieger, Jürgen Glas, Talin Haritunians, Ailsa Hart, Chris Hawkey, Matija Hedl, Xinli Hu, Tom H Karlsen, Limas Kupcinskas, Subra Kugathasan, Anna Latiano, Debby Laukens, Ian C Lawrance, Charlie W Lees, Edouard Louis, Gillian Mahy, John Mansfield, Angharad R Morgan, Craig Mowat, William Newman, Orazio Palmieri, Cyriel Y Ponsioen, Uros Potocnik, Natalie J Prescott, Miguel Regueiro, Jerome I Rotter, Richard K Russell, Jeremy D Sanderson, Miquel Sans, Jack Satsangi, Stefan Schreiber, Lisa A Simms, Jurgita Sventoraityte, Stephan R Targan, Kent D Taylor, Mark Tremelling, Hein W Verspaget, Martine De Vos, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Sebastian Zeissig, Bin Zhang, Clarence K Zhang, Hongyu Zhao, International IBD Genetics Consortium (IIBDGC), Mark S Silverberg, Vito Annese, Hakon Hakonarson, Steven R Brant, Graham Radford-Smith, Christopher G Mathew, John D Rioux, Eric E Schadt, Mark J Daly, Andre Franke, Miles Parkes, Séverine Vermeire, Jeffrey C Barrett, and Judy H Cho. **Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.** *Nature*, **491**(7422):119–124, November 2012. 64, 137, 196, 199, 200

[280] P J Morrissey, K Charrier, S Braddy, D Liggitt, and J D Watson. **CD4+ T cells that express high levels of CD45RB induce wasting disease when transferred into congenic severe combined immunodeficient mice. Disease development is prevented by co-transfer of purified CD4+ T cells.** *The Journal of experimental medicine*, **178**(1):237–244, July 1993. 64

[281] M F Neurath. **The Transcription Factor T-bet Regulates Mucosal T Cell Activation in Experimental Colitis and Crohn's Disease**. *Journal of Experimental Medicine*, **195**(9):1129–1143, May 2002. 64

[282] Ken Cadwell, Khushbu K Patel, Nicole S Maloney, Ta-Chiang Liu, Aylwin C Y Ng, Chad E Storer, Richard D Head, Ramnik Xavier, Thaddeus S Stappenbeck, and Herbert W Virgin. **Virus-Plus-Susceptibility Gene Interaction Determines Crohn's Disease Gene Atg16L1 Phenotypes in Intestine**. *Cell*, **141**(7):1135–1145, June 2010. 64

[283] W Dieterich, T Ehnis, M Bauer, P Donner, U Volta, E O Riecken, and D Schuppan. **Identification of tissue transglutaminase as the autoantigen of celiac disease.** *Nature Medicine*, **3**(7):797–801, July 1997. 65

[284] David A van Heel, Lude Franke, Karen A Hunt, Rhian Gwilliam, Alexandra Zhernakova, Mike Inouye, Martin C Wapenaar, Martin C N M Barnardo, Graeme Bethel, Geoffrey K T Holmes, Con Feighery, Derek Jewell, Dermot Kelleher, Parveen Kumar, Simon Travis, Julian RF Walters, David S Sanders, Peter Howdle, Jill Swift, Raymond J Playford, William M McLaren, M Luisa Mearin, Chris J Mulder, Ross McManus, Ralph McGinnis, Lon R Cardon, Panos Deloukas, and Cisca Wijmenga. **A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21**. *Nature Genetics*, **39**(7):827–829, June 2007. 65

[285] P Collin, J Salmi, O Hällström, H Oksa, H Oksala, M Mäki, and T Reunala. **High frequency of coeliac disease in adult patients with type-I diabetes.** *Scandinavian journal of gastroenterology*, **24**(1):81–84, January 1989. 65

[286] Gosia Trynka, Cisca Wijmenga, and David A van Heel. **A genetic perspective on coeliac disease**. *Trends in Molecular Medicine*, **16**(11):537–550, November 2010. 65

[287] L Greco, R Romino, I Coto, N Di Cosmo, S Percopo, M Maglio, F Paparo, V Gasperi, M G Limongelli, R Cotichini, C D'Agate, N Tinto, L Sacchetti, R Tosi, and M A Stazi. **The first large population based twin study of coeliac disease.** *Gut*, **50**(5):624–628, May 2002. 65

[288] Y Claire Wang, Klim McPherson, Tim Marsh, Steven L Gortmaker, and Martin Brown. **Health and economic burden of the projected obesity trends in the USA and the UK.** *Lancet*, **378**(9793):815–825, August 2011. 66

[289] Steven L Gortmaker, Boyd A Swinburn, David Levy, Rob Carter, Patricia L Mabry, Diane T Finegood, Terry Huang, Tim Marsh, and Marjory L Moodie. **Changing the future of obesity: science, policy, and action.** *Lancet*, **378**(9793):838–847, August 2011. 66

[290] Y Zhang, R Proenca, M Maffei, M Barone, L Leopold, and J M Friedman. **Positional cloning of the mouse obese gene and its human homologue.** *Nature*, **372**(6505):425–432, December 1994. 66

[291] L A Campfield, F J Smith, Y Guisez, R Devos, and P Burn. **Recombinant mouse OB protein: evidence for a peripheral signal linking adiposity and central neural networks.** *Science*, **269**(5223):546–549, July 1995. 66

[292] C T Montague, I S Farooqi, J P Whitehead, M A Soos, H Rau, N J Wareham, C P Sewter, J E Digby, S N Mohammed, J A Hurst, C H Cheetham, A R Earley, A H Barnett, J B Prins, and S O'Rahilly. **Congenital leptin deficiency is associated with severe early-onset obesity in humans.** *Nature*, **387**(6636):903–908, June 1997. 66

[293] S B Cassidy, E Dykens, and C A Williams. **Prader-Willi and Angelman syndromes: sister imprinted disorders.** *American journal of medical genetics*, **97**(2):136–146, 2000. 66

[294] T M Frayling, N J Timpson, M N Weedon, E Zeggini, R M Freathy, C M Lindgren, J R B Perry, K S Elliott, H Lango, N W Rayner, B Shields, L W Harries, J C Barrett, S Ellard, C J Groves, B Knight, A M Patch, A R Ness, S Ebrahim, D A Lawlor, S M Ring, Y Ben-Shlomo, M R Jarvelin, U Sovio, A J Bennett, D Melzer, L Ferrucci, R J F Loos, I Barroso, N J Wareham, F Karpe, K R Owen, L R Cardon, M Walker, G A Hitman, C N A Palmer, A S F Doney, A D Morris, G D Smith, The Wellcome Trust Case Control Consortium, A T Hattersley, and M I McCarthy. **A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity**. *Science*, **316**(5826):889–894, May 2007. 66

[295] David Meyre, Jérôme Delplanque, Jean-Claude Chèvre, Cecile Lecoeur, Stéphane Lobbens, Sophie Gallina, Emmanuelle Durand, Vincent Vatin, Franck Degraeve, Christine Proença, Stefan Gaget, Antje Körner, Peter Kovacs, Wieland Kiess, Jean Tichet, Michel Marre, Anna-Liisa Hartikainen, Fritz Horber, Natascha Potoczna, Serge Hercberg, Claire Levy-Marchal, François Pattou, Barbara Heude, Maithé Tauber, Mark I McCarthy, Alexandra I F Blakemore, Alexandre Montpetit, Constantin Polychronakos, Jacques Weill, Lachlan J M Coin, Julian Asher, Paul Elliott, Marjo-Riitta Jarvelin, Sophie Visvikis-Siest, Beverley Balkau, Rob Sladek, David Balding, Andrew Walley, Christian Dina, and Philippe Froguel. **Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations**. *Nature Genetics*, **41**(2):157–159, January 2009. 66

[296] Tove Fall and Erik Ingelsson. **Genome-wide association studies of obesity and metabolic syndrome**. *Molecular and Cellular Endocrinology*, September 2012. 66

[297] Peter J Turnbaugh, Fredrik Bäckhed, Lucinda Fulton, and Jeffrey I Gordon. **Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome**. *Cell Host & Microbe*, **3**(4):213–223, April 2008. 67

[298] M Vijay-Kumar, J D Aitken, F A Carvalho, T C Cullender, S Mwangi, S Srinivasan, S V Sitaraman, R Knight, R E Ley, and A T Gewirtz. **Metabolic Syndrome and Altered Gut Microbiota in Mice Lacking Toll-Like Receptor 5**. *Science*, **328**(5975):228–231, April 2010. 67

[299] Carey N Lumeng, Ivan Maillard, and Alan R Saltiel. **T-ing up inflammation in fat.** *Nature Medicine*, **15**(8):846–847, August 2009. 67

[300] David Clayton and Hin-Tak Leung. **An R Package for Analysis of Whole-Genome Association Studies**. *Human Heredity*, **64**(1):45–51, 2007. 69, 94

[301] Tao Liu, Jorge A Ortiz, Len Taing, Clifford A Meyer, Bernett Lee, Yong Zhang, Hyunjin Shin, Swee S Wong, Jian Ma, Ying Lei, Utz J Pape, Michael Poidinger, Yiwen Chen, Kevin Yeung, Myles Brown, Yaron Turpaz, and X Shirley Liu. **Cistrome: an integrative platform for transcriptional regulation studies.** *Genome biology*, **12**(8):R83, 2011. 70

[302] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. **Model-based analysis of ChIP-Seq (MACS).** *Genome biology*, **9**(9):R137, 2008. 70, 96

[303] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. **The accessible chromatin landscape of the human genome**. *Nature*, **489**(7414):75–82, September 2012. 70, 179

[304] M T Maurano, R Humbert, E Rynes, R E Thurman, E Haugen, H Wang, A P Reynolds, R Sandstrom, H Qu, J Brody, A Shafer, F Neri, K Lee, T Kutyavin, S Stehling-Sun, A K Johnson, T K Canfield, E Giste, M Diegel, D Bates, R S Hansen, S Neph, P J Sabo, S Heimfeld, A Raubitschek, S Ziegler, C Cotsapas, N Sotoodehnia, I Glass, S R Sunyaev, R Kaul, and J A Stamatoyannopoulos. **Systematic Localization of Common Disease-Associated Variation in Regulatory DNA**. *Science*, **337**(6099):1190–1195, September 2012. 70, 179

[305] Charles E Grant, Timothy L Bailey, and William Stafford Noble. **FIMO: scanning for occurrences of a given motif.** *Bioinformatics*, **27**(7):1017–1018, April 2011. 71, 103

[306] T L Bailey, M Boden, F A Buske, M Frith, C E Grant, L Clementi, J Ren, W W Li, and W S Noble. **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Research*, **37**(Web Server):W202–W208, June 2009. 71, 103, 104

[307] Morgane Thomas-Chollier, Andrew Hufton, Matthias Heinig, Sean O'Keeffe, Nassim El Masri, Helge G Roider, Thomas Manke, and Martin Vingron. **Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs**. *Nature Protocols*, **6**(12):1860–1869, November 2011. 71, 109

[308] Wei Liao, Jian-Xin Lin, Lu Wang, Peng Li, and Warren J Leonard. **Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages**. *Nature Immunology*, **12**(6):551–559, April 2011. 73

[309] J D Stender, K Kim, T H Charn, B Komm, K C N Chang, W L Kraus, C Benner, C K Glass, and B S Katzenellenbogen. **Genome-Wide Analysis of Estrogen Receptor DNA Binding and Tethering Mechanisms Identifies Runx1 as a Novel Tethering Factor in Receptor-Mediated Transcriptional Activation**. *Molecular and Cellular Biology*, **30**(16):3943–3955, July 2010. 73

[310] Refik Gökmen. *The role of the transcription factor T-bet in the differentiation and function of T helper cell subsets*. PhD thesis, School of Medicine, Kings College London, 2010. 79

[311] Kerem Atalar. *The role of microRNAs in T helper cell differentiation and function*. PhD thesis, Kings College London, 2012. 79

[312] Karin L Heckman and Larry R Pease. **Gene splicing and mutagenesis by PCR-driven overlap extension**. *Nature Protocols*, **2**(4):924–932, April 2007. 86

[313] Yuka Kanno, Golnaz Vahedi, Kiyoshi Hirahara, Kentner Singleton, and John J O'Shea. **Transcriptional and Epigenetic Control of T Helper Cell Specification: Molecular Mechanisms Underlying Commitment and Plasticity ***. *Annual review of immunology*, **30**(1):707–731, April 2012. 95

[314] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature*, **473**(7345):43–49, March 2011. 97

[315] P Wang, M Dai, W Xuan, R C McEachin, A U Jackson, L J Scott, B Athey, S J Watson, and F Meng. **SNP Function Portal: a web database for exploring the function implication of SNP alleles**. *Bioinformatics*, **22**(14):e523–e529, July 2006. 98

[316] P Rice, I Longden, and A Bleasby. **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends in genetics : TIG*, **16**(6):276–277, June 2000. 99

[317] R Jothi, S Cuddapah, A Barski, K Cui, and K Zhao. **Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data**. *Nucleic Acids Research*, **36**(16):5221–5231, August 2008. 116

[318] A S Brodsky. **A Microbead-based System for Identifying and Characterizing RNA-Protein Interactions by Flow Cytometry**. *Molecular & Cellular Proteomics*, **1**(12):922–929, December 2002. 123

[319] Andre Franke, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A Anderson, Joshua C Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G Mathew, Grant W Montgomery, Natalie J Prescott, Soumya Raychaudhuri, Jerome I Rotter, Philip Schumm, Yashoda Sharma, Lisa A Simms, Kent D Taylor, David Whiteman, Cisca Wijmenga, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Büning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jürgen Glas, Andre Van Gossum, Stephen L Guthery, Jonas Halfvarson, Hein W Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lémann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julián Panés, Anne Phillips, Deborah D Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A Hillary Steinhart, Pieter C F Stokkers, Leif Törkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R Targan, Steven R Brant, John D Rioux, Mauro D'Amato, Rinse K Weersma, Subra Kugathasan, Anne M Griffiths, John C Mansfield, Séverine Vermeire, Richard H Duerr, Mark S Silverberg, Jack Satsangi, Stefan Schreiber, Judy H Cho, Vito Annese, Hakon Hakonarson, Mark J Daly, and Miles Parkes. **Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci**. *Nature Genetics*, **42**(12):1118–1125, November 2010. 137, 139

[320] Dara G Torgerson, Elizabeth J Ampleford, Grace Y Chiu, W James Gauderman, Christopher R Gignoux, Penelope E Graves, Blanca E Himes, Albert M Levin, Rasika A Mathias, Dana B Hancock, James W Baurley, Celeste Eng, Debra A Stern, Juan C Celedón, Nicholas Rafaels, Daniel Capurso, David V Conti, Lindsey A Roth, Manuel Soto-Quiros, Alkis Togias, Xingnan Li, Rachel A Myers, Isabelle Romieu, David J Van Den Berg, Donglei Hu, Nadia N Hansel, Ryan D Hernandez, Elliott Israel, Muhammad T Salam, Joshua Galanter, Pedro C Avila, Lydiana Avila, Jose R Rodriquez-Santana, Rocio Chapela, William Rodriguez-Cintron, Gregory B Diette, N Franklin Adkinson, Rebekah A Abel, Kevin D Ross, Min Shi, Mezbah U Faruque, Georgia M Dunston, Harold R Watson, Vito J Mantese, Serpil C Ezurum, Liming Liang, Ingo Ruczinski, Jean G Ford, Scott Huntsman, Kian Fan Chung, Hita Vora, Xia Li, William J Calhoun, Mario Castro, Juan J Sienra-Monge, Blanca del Rio-Navarro, Klaus A Deichmann, Andrea Heinzmann, Sally E Wenzel, William W Busse, James E Gern, Robert F Lemanske, Terri H Beaty, Eugene R Bleecker, Benjamin A Raby, Deborah A Meyers, Stephanie J London, Frank D Gilliland, Esteban G Burchard, Fernando D Martinez, Scott T Weiss, L Keoki Williams, Kathleen C Barnes, Carole Ober, and Dan L Nicolae. **Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations**. *Nature Genetics*, **43**(9):887–892, July 2011. 137

[321] Yukinori Okada, Tomomitsu Hirota, Yoichiro Kamatani, Atsushi Takahashi, Hiroko Ohmiya, Natsuhiko Kumasaka, Koichiro Higasa, Yumi Yamaguchi-Kabata, Naoya Hosono, Michael A Nalls, Ming Huei Chen, Frank J A van Rooij, Albert V Smith, Toshiko Tanaka, David J Couper, Neil A Zakai, Luigi Ferrucci, Dan L Longo, Dena G Hernandez, Jacqueline C M Witteman, Tamara B Harris, Christopher J O'Donnell, Santhi K Ganesh, Koichi Matsuda, Tatsuhiko Tsunoda, Toshihiro Tanaka, Michiaki Kubo, Yusuke Nakamura, Mayumi Tamari, Kazuhiko Yamamoto, and Naoyuki Kamatani. **Identification of Nine Novel Loci Associated with White Blood Cell Subtypes in a Japanese Population**. *PLoS Genetics*, **7**(6):e1002067, June 2011. 137, 200

[322] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez Arcelus, Alkes Price, Towfique Raj, James Nisbett, Alexandra C Nica, Claude Beazley, Richard Durbin, Panos Deloukas, and Emmanouil T Dermitzakis. **Patterns of Cis Regulatory Variation in Diverse Human Populations**. *PLoS Genetics*, **8**(4):e1002639, April 2012. 139

[323] Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, Åsa K Hedman, Veronique Bataille, Jordana Tzenova Bell, Gabriela Surdulescu, Antigone S Dimas, Catherine Ingle, Frank O Nestle, Paola di Meglio, Josine L Min, Alicja Wilk, Christopher J Hammond, Neelam Hassanali, Tsun-Po Yang, Stephen B Montgomery, Steve O'Rahilly, Cecilia M Lindgren, Krina T Zondervan, Nicole Soranzo, Inês Barroso, Richard Durbin, Kourosh Ahmadi, Panos Deloukas, Mark I McCarthy, Emmanouil T Dermitzakis, Timothy D Spector, and The MuTHER Consortium. **The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study**. *PLoS Genetics*, **7**(2):e1002003, February 2011. 139

[324] M D'Agostino, G Tornillo, M G Caporaso, M V Barone, E Ghigo, S Bonatti, and G Mottola. **Ligand of Numb proteins LNX1p80 and LNX2 interact with the human glycoprotein CD8 and promote its ubiquitylation and endocytosis**. *Journal of Cell Science*, **124**(21):3545–3556, November 2011. 141

[325] Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan, Reedik Mägi, Joshua C Randall, Sailaja Vedantam, Thomas W Winkler, Lu Qi, Tsegaselassie Workalemahu, Iris M Heid, Valgerdur Steinthorsdottir, Heather M Stringham, Michael N Weedon, Eleanor Wheeler, Andrew R Wood, Teresa Ferreira, Robert J Weyant, Ayellet V Segrè, Karol Estrada, Liming Liang, James Nemesh, Ju-Hyun Park, Stefan Gustafsson, Tuomas O Kilpeläinen, Jian Yang, Nabila Bouatia-Naji, Tõnu Esko, Mary F Feitosa, Zoltán Kutalik, Massimo Mangino, Soumya Raychaudhuri, Andre Scherag, Albert Vernon Smith, Ryan Welch, Jing Hua Zhao, Katja K Aben, Devin M Absher, Najaf Amin, Anna L Dixon, Eva Fisher, Nicole L Glazer, Michael E Goddard, Nancy L Heard-Costa, Volker Hoesel, Jouke-Jan Hottenga, Åsa Johansson, Toby Johnson, Shamika Ketkar, Claudia Lamina, Shengxu Li, Miriam F Moffatt, Richard H Myers, Narisu Narisu, John R B Perry, Marjolein J Peters, Michael Preuss, Samuli Ripatti, Fernando Rivadeneira, Camilla Sandholt, Laura J Scott, Nicholas J Timpson, Jonathan P Tyrer, Sophie van Wingerden, Richard M Watanabe, Charles C White, Fredrik Wiklund, Christina Barlassina, Daniel I Chasman, Matthew N Cooper, John-Olov Jansson, Robert W Lawrence, Niina Pellikka, Inga Prokopenko, Jianxin Shi, Elisabeth Thiering, Helene Alavere, Maria T S Alibrandi, Peter Almgren, Alice M Arnold, Thor Aspelund, Larry D Atwood, Beverley Balkau, Anthony J Balmforth, Amanda J Bennett, Yoav Ben-Shlomo, Richard N Bergman, Sven Bergmann, Heike Biebermann, Alexandra I F Blakemore, Tanja Boes, Lori L Bonnycastle, Stefan R Bornstein, Morris J Brown, Thomas A Buchanan, Fabio Busonero, Harry Campbell, Francesco P Cappuccio, Christine Cavalcanti-Proença, Yii-Der Ida Chen, Chih-Mei Chen, Peter S Chines, Robert Clarke, Lachlan Coin, John Connell, Ian N M Day, Martin den Heijer, Jubao Duan, Shah Ebrahim, Paul Elliott, Roberto Elosua, Gudny Eiriksdottir, Michael R Erdos, Johan G Eriksson, Maurizio F Facheris, Stephan B Felix, Pamela Fischer-Posovszky, Aaron R Folsom, Nele Friedrich, Nelson B Freimer, Mao Fu, Stefan Gaget, Pablo V Gejman, Eco J C Geus, Christian Gieger, Anette P Gjesing, Anuj Goel, Philippe Goyette, Harald Grallert, Jürgen Gräßler, Danielle M Greenawalt, Christopher J Groves, Vilmundur Gudnason, Candace Guiducci, Anna-Liisa Hartikainen, Neelam Hassanali, Alistair S Hall, Aki S Havulinna, Caroline Hayward, Andrew C Heath, Christian Hengstenberg, Andrew A Hicks, Anke Hinney, Albert Hofman, Georg Homuth, Jennie Hui, Wilmar Igl, Carlos Iribarren, Bo Isomaa, Kevin B Jacobs, Ivonne Jarick, Elizabeth Jewell, Ulrich John, Torben Jørgensen, Pekka Jousilahti, Antti Jula, Marika Kaakinen, Eero Kajantie, Lee M Kaplan, Sekar Kathiresan, Johannes Kettunen, Leena Kinnunen, Joshua W Knowles, Ivana Kolcic, Inke R König, Seppo Koskinen, Peter Kovacs, Johanna Kuusisto, Peter Kraft, Kirsti Kvaløy, Jaana Laitinen, Olivier Lantieri, Chiara Lanzani, Lenore J Launer, Cecile Lecoeur, Terho Lehtimäki, Guillaume Lettre, Jianjun Liu, Marja-Liisa Lokki, Mattias Lorentzon, Robert N Luben, Barbara Ludwig, Paolo Manunta, Diana Marek, Michel Marre, Nicholas G Martin, Wendy L McArdle, Anne McCarthy, Barbara McKnight, Thomas Meitinger, Olle Melander, David Meyre, Kristian Midthjell, Grant W Montgomery, Mario A Morken, Andrew P Morris, Rosanda Mulic, Julius S Ngwa, Mari Nelis, Matt J Neville, Dale R Nyholt, Christopher J O'Donnell, Stephen O'Rahilly, Ken K Ong, Ben Oostra, Guillaume Paré, Alex N Parker, Markus Perola, Irene Pichler, Kirsi H Pietiläinen, Carl G P Platou, Ozren Polasek, Anneli Pouta, Suzanne Rafelt, Olli Raitakari, Nigel W Rayner, Martin Ridderstråle, Winfried Rief, Aimo Ruokonen, Neil R Robertson, and ... Rzehak. **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nature Genetics*, **42**(11):937–948, October 2010. 141

[326] G F Gerberick, L W Cruse, C M Miller, E E Sikorski, and G M Ridder. **Selective modulation of T cell memory markers CD62L and CD44 on murine draining lymph node cells following allergen and irritant treatment.** *Toxicology and applied pharmacology*, **146**(1):1–10, September 1997. 148

[327] T L Freitag, S Rietdijk, Y Junker, Y Popov, A K Bhan, C P Kelly, C Terhorst, and D Schuppan. **Gliadin-primed CD4+CD45RBlowCD25- T cells drive gluten-dependent small intestinal damage after adoptive transfer into lymphopenic mice**. *Gut*, **58**(12):1597–1605, November 2009. 159, 164

[328] Shohei Hori and Shimon Sakaguchi. **Foxp3: a critical regulator of the development and function of regulatory T cells**. *Microbes and Infection*, **6**(8):745–751, July 2004. 170

[329] 1000 Genomes Project Consortium, Gonçalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. **An integrated map of genetic variation from 1,092 human genomes.** *Nature*, **491**(7422):56–65, November 2012. 177, 194, 195

[330] M W Steinberg, O Turovskaya, R B Shaikh, G Kim, D F McCole, K Pfeffer, K M Murphy, C F Ware, and M Kronenberg. **A crucial role for HVEM and BTLA in preventing intestinal inflammation**. *Journal of Experimental Medicine*, **205**(6):1463–1476, June 2008. 184

[331] R L Modlin. **Th1-Th2 paradigm: insights from leprosy.** *The Journal of investigative dermatology*, **102**(6):828–832, June 1994. 185

[332] Kyu-Yeob Kim, Hyun-Ja Jeong, and Hyung-Min Kim. **The role of T-bet in obesity: lack of T-bet causes obesity in male mice**. *The Journal of Nutritional Biochemistry*, August 2012. 186, 198

[333] Stephan Armin Wacker, Cristobal Alvarado, Götz von Wichert, Uwe Knippschild, Jörg Wiedenmann, Karen Clauss, Gerd Ulrich Nienhaus, Horst Hameister, Bernd Baumann, Tilman Borggrefe, Walter Knöchel, and Franz Oswald. **RITA, a novel modulator of Notch signalling, acts via nuclear export of RBP-J.** *The EMBO journal*, **30**(1):43–56, January 2011. 187

7

Appendix 1

7.1 Script for Determining Binding Site SNPs

```
library(snpMatrix)
my.chroms <- c(1:22)
window <- 0
min.r2 <- 0.8
TransFact.results.snptoo <- numeric()
TransFact.results.snp <- numeric()
TransFact.results.linkage <- numeric()
absent <- numeric()


system("awk '{print $1,$11,$12}' /Users/claresoderquest/Documents/katrina/AllStudies
Analysis/041111DatabaseWork/Gata3BindingSitesTh2Cells.txt > TransFact_katrina.tab")

TF <- read.table('TransFact_katrina.tab', header = TRUE)

for (chrom in my.chroms) {
cat("Looking at chromosome", chrom, "\n")

my.table <- paste('/Users/claresoderquest/Documents/katrina/AllStudiesAnalysis
/041111DatabaseWork/FullGWASTableForScript.txt')

if (file.exists(my.table)) {

Chroms <- read.table ( file = my.table, header = TRUE, row.names = 1)

ChromsCurrent <- subset(Chroms, (Chr == chrom))

hapmap.data <- paste('file:////Users/claresoderquest/Documents/katrina/HapMap3/chr',
chrom, 'CEU.txt', sep ='')
```

```
hapmap <- read.HapMap.data( url = hapmap.data, verbose = FALSE, save = NULL )

for (j in 1:nrow(ChromsCurrent)) {
snp <- row.names(ChromsCurrent)[j]
if(!(snp %in% dimnames(hapmap$snp.data)[[2]])) {
cat("Cannot find snp", snp, "in hapmap\n")
absent <- c(absent, snp)
} else {
my.ld.results <- ld.with(data = hapmap$snp.data, snps = snp)$rsq2
high.ld.snps <- subset(row.names(my.ld.results), (my.ld.results[, snp] > min.r2)
& !is.na(my.ld.results[, snp]))
my.snps.support <- hapmap$snp.support[ high.ld.snps,]


first <- TRUE

for (i in 1:nrow(my.snps.support)) {

snptoo <- row.names(my.snps.support)[i]
position <- my.snps.support$Position[i]
TableChrom <- paste("chr", chrom, sep = ”)


# TF
TF.intersec <- subset(TF, (Chr == TableChrom) & (position > rStart - window) &
(position < rEnd + window) )
if (nrow(TF.intersec) >= 1) {
cat("Looking at current TF snp", snptoo, "in LD with snp ", snp, "\n")
print(TF.intersec)
if (first) {
print(subset(my.ld.results, (my.ld.results[, snp] > min.r2) & !is.na(my.ld.results[, snp])))
first <- FALSE}
linkage.value.TF <- my.ld.results[snptoo, snp]
print(linkage.value.TF)
TransFact.results.snptoo <- c(TransFact.results.snptoo, snptoo)
TransFact.results.snp <- c(TransFact.results.snp, snp)
TransFact.results.linkage <- c(TransFact.results.linkage, linkage.value.TF)}


TF.results <- cbind(TransFact.results.snptoo, TransFact.results.snp, TransFact.results.linkage)


} } }}}

write.table(TF.results, "/Users/claresoderquest/Documents/katrina/AllStudiesAnalysis
/041111Database Work/ResultsGata3BindingSitesTh2Cells.txt", row.names = FALSE, quote
= FALSE, sep = "\t")
```

write.table(absent, "/Users/claresoderquest/Documents/katrina/AllStudiesAnalysis /041111Database Work/AbsentResultsGata3BindingSitesTh2Cells.txt", row.names = FALSE, quote = FALSE, sep = "\t")

7.2 Lists of Strictly and Loosely Immune Mediated Conditions

List of NHGRI GWAS catalogue traits judged as definitely immune mediated:

Ankylosing spondylitis
Arthritis (juvenile idiopathic)
Asthma (aspirin-intolerant)
Asthma (childhood onset)
Asthma
Atopic dermatitis
Atopy
Behcet's disease
Coeliac disease and Rheumatoid arthritis
Coeliac disease
Crohn's disease and Coeliac disease
Crohn's disease and sarcoidosis (combined)
Crohn's disease
Eosinophilic esophagitis (pediatric)
Graves' disease
Hepatitis B vaccine response
Inflammatory bowel disease (early onset)
Inflammatory bowel disease
Lupus
Multiple sclerosis (age of onset)
Multiple sclerosis (severity)
Multiple sclerosis
Neonatal lupus
Primary biliary cirrhosis
Psoriasis
Psoriatic arthritis

Rheumatoid arthritis
Sarcoidosis
Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
Stevens-Johnson syndrome
Systemic lupus erythematosus
Systemic sclerosis
Type 1 diabetes autoantibodies
Type 1 diabetes
Ulcerative colitis
Vitiligo

List of NHGRI GWAS catalogue traits judged as possibly immune mediated:

Acute lymphoblastic leukaemia (childhood)
Age-related macular degeneration (wet)
AIDS
AIDS progression
Alopecia areata
Amyloid A Levels
Anti-cyclic Citrullinated Peptide Antibody
Asthma (toluene diisocyanate-induced)
Brain lesion load
Breast cancer (prognosis)
C-reactive protein
Carotid atherosclerosis in HIV infection
Carotid intima media thickness
CD4:CD8 lymphocyte ratio
Chronic Hepatitis C infection
Chronic kidney disease
Chronic lymphocytic leukaemia
Colorectal cancer
Common variable immunodeficiency
Diabetic retinopathy
Drug-induced liver injury (amoxicillin-clavulanate)
Drug-induced liver injury (flucloxacillin)

End-stage renal disease
End-stage renal disease (non-diabetic)
Epirubicin-induced leukopenia
Erythrocyte sedimentation rate
Follicular lymphoma
Glomerulosclerosis
Hepatitis B
Hepatocellular carcinoma
HIV (mother-to-child transmission)
HIV-1 control
HIV-1 progression
HIV-1 replication
HIV-1 susceptibility
HIV-1 viral setpoint
Hodgkin's lymphoma
Idiopathic pulmonary fibrosis
IFN-related cytopenia
Immunoglobulin A
Interleukin-18 levels
Interstitial lung disease
Kawasaki disease
Keloid
Lapatinib-induced hepatotoxicity
Large B-cell lymphoma
Leprosy
Malaria
Meningococcal disease
Monocyte early outgrowth colony forming units
Moyamoya disease
Multiple sclerosis–Brain Glutamate Levels
Myeloproliferative neoplasms Nasopharyngeal carcinoma
Nephropathy (idiopathic membranous)
Nephropathy
Nephrotic syndrome (acquired)
Neuromyelitis optica
Neutrophil count

Non-alcoholic fatty liver disease histology (AST)
Non-alcoholic fatty liver disease histology (lobular)
Non-alcoholic fatty liver disease histology (other)
Nonalcoholic fatty liver disease
Periodontitis
Plasma C4b binding protein levels
Plasma eosinophil count
Premature ovarian failure
Primary sclerosing cholangitis
Renal function and chronic kidney disease
Response to hepatitis C treatment
Response to interferon beta therapy
Response to TNF antagonist treatment
Response to treatment for acute lymphoblastic leukaemia
Serologic markers in systemic lupus erythematosus
Serum IgE levels
Serum matrix metalloproteinase
Serum soluble E-selectin
Soluble ICAM-1
Soluble leptin receptor levels
Soluble levels of adhesion molecules
T-tau
Tuberculosis
Vaccine-related adverse events
White blood cell count
White blood cell types

List of NHGRI GWAS traits judged as non immune mediated:

5-HTT brain serotonin transporter levels
AB1-42
Abdominal aortic aneurysm
Acenocoumarol maintenance dosage
Activated partial thromboplastin time
Addiction
Adiponectin levels

Adiposity
Adverse response to aromatase inhibitors
Adverse response to carbamapezine
Age-related macular degeneration
Aging
Aging (time to death)
Aging (time to event)
Aging traits
Alcohol consumption
Alcohol dependence
Alcoholism (12-month weekly alcohol consumption)
Alcoholism (alcohol dependence factor score)
Alcoholism (alcohol use disorder factor score)
Alcoholism (heaviness of drinking)
Alzheimer's disease
Alzheimer's disease (late onset)
Alzheimer's disease biomarkers
Amyotrophic lateral sclerosis
Amyotrophic lateral sclerosis (interaction)
Angiotensin-converting enzyme activity
Anorexia nervosa
Anthropometric traits
Antipsychotic-induced QTc interval prolongation
Aortic root size
Arterial stiffness
Aspartate aminotransferase
Asperger disorder
Atrial fibrillation
Atrial fibrillation/atrial flutter
Atrioventricular conduction
Attention deficit hyperactivity disorder
Attention deficit hyperactivity disorder (time to onset)
Attention deficit hyperactivity disorder and conduct disorder
Attention deficit hyperactivity disorder motor coordination
Attention deficit hyperactivity disorder symptoms (interaction)
Autism

Basal cell carcinoma (cutaneous)
Beta thalassemia/haemoglobin E disease
Biliary atresia
Bilirubin levels
Biochemical measures
Biomedical quantitative traits
Bipolar disorder
Bipolar disorder (age of onset and psychomotor symptoms)
Bipolar disorder and major depressive disorder (combined)
Bipolar disorder and schizophrenia
Birth weight
Bitter taste response
Black vs. blond hair colour
Black vs. red hair colour
Bladder cancer
Bleomycin sensitivity
Blond vs. brown hair colour
Blood lipid traits
Blood pressure
Blue vs. brown eyes
Blue vs. green eyes
Body mass (lean)
Body mass in chronic obstructive pulmonary disease
Body mass index
Body mass index and fat mass
Bone mineral density
Bone mineral density (hip)
Bone mineral density (spine)
Brain imaging
Brain imaging in schizophrenia (interaction)
Brain structure
Breast cancer
Bronchopulmonary dysplasia
Burning and freckling
Caffeine consumption
Cannabis dependence

Cardiac hypertrophy
Cardiac structure and function
Cardiovascular disease risk factors
Cataracts in type 2 diabetes
Caudate nucleus volume
Central corneal thickness
Chemotherapeutic susceptibility
Cholesterol
Cholesterol, total
Chronic fatigue syndrome
Chronic kidney disease and serum creatinine levels
Chronic myeloid leukaemia
Chronic obstructive pulmonary disease
Cleft lip
Coffee consumption
Cognitive ability
Cognitive performance
Cognitive test performance
Common traits (Other)
Conduct disorder (case status)
Conduct disorder (interaction)
Conduct disorder (symptom count)
Corneal curvature
Corneal structure
Coronary artery calcification
Coronary heart disease
Coronary restenosis
Coronary spasm
Cortisol secretion
Creutzfeldt-Jakob disease
Cutaneous nevi
Cystatin C
Cystic fibrosis severity
D-dimer levels
Dental caries
Depression–quantitative trait

Diabetes (incident)
Diabetes related insulin traits
Diabetic nephropathy
Dialysis-related mortality
Diastolic blood pressure
Digit length ratio
Dilated cardiomyopathy
Drinking behaviour
Dupuytren's disease
Echocardiographic traits
Educational attainment
Electrocardiographic conduction measures
Electrocardiographic traits
Electroencephalogram traits
Emphysema-related traits
Endometrial cancer
Endometriosis
Endothelial function traits
Entorhinal cortical thickness
Environmental confusion in the home
Episodic memory
Erectile dysfunction and prostate cancer treatment
Essential tremor
Event-related brain oscillations
Exercise (leisure time)
Exercise treadmill test traits
Eye colour
Eye color traits
F-cell distribution
Factor VII
Fasting glucose-related traits
Fasting insulin-related traits
Fasting plasma glucose
Femoral neck bone geometry
Foetal haemoglobin levels
Fibrinogen

Folate pathway vitamin levels
Fracture-related traits
Freckles
Freckling
Frontotemporal lobar degeneration
Fuchs's corneal dystrophy
Functional MRI
Gallstones
General cognitive ability
Glaucoma
Glaucoma (exfoliation)
Glaucoma (primary open-angle)
Glioma
Glioma (high-grade)
Glycated haemoglobin levels
Glycemic control in type 1 diabetes (HbA1c)
Haematological and biochemical traits
Haematological parameters
Haematology traits
Haemoglobin
Haemoglobin levels
Hair color
Hair morphology
Handedness in dyslexia
HDL cholesterol
HDL Cholesterol - Triglycerides (HDLC-TG)
Hearing impairment
Heart failure
Heart rate variability traits
Height
Hematocrit
Hemostatic factors and haematological phenotypes
Heroin addiction
Hip bone size
Hip geometry
Hippocampal atrophy

Hirschsprung's disease
Hoarding
Homocysteine levels
HPV seropositivity
Hyperactive-impulsive symptoms
Hypertension
Hypertension (young onset)
Hypertriglyceridemia
Hypospadias
Ileal carcinoids
Inattentive symptoms
Information processing speed
Insulin resistance/response
Insulin traits
Insulin-like growth factors
Intelligence
Intracranial aneurysm
Iris characteristics
Iris color
Ischemic stroke
Kidney stones
Knee osteoarthritis
LDL cholesterol
Left ventricular mass
Lipoprotein-associated phospholipase A2 activity and mass
Longevity
Lumiracoxib-related liver injury
Lung adenocarcinoma
Lung adenocarcinoma (clinical stage)
Lung cancer
Major CVD
Major depressive disorder
Major depressive disorder (broad)
Major mood disorders
Male infertility
Male-pattern baldness

Mammographic density
Mathematical ability
Mean corpuscular haemoglobin
Mean corpuscular volume
Mean forced vital capacity from 2 exams
Mean platelet volume
Melanoma
Memory (short-term)
Memory performance
Menarche (age at onset)
Menarche and menopause (age at onset)
Meningioma
Menopause (age at onset)
Metabolic syndrome
Metabolic syndrome (bivariate traits)
Metabolic traits
Methamphetamine dependence
Migraine
Migraine in bipolar disorder
Morbidity-free survival
Mortality among heart failure patients
MRI atrophy measures
Myocardial infarction
Myocardial infarction (early onset)
Myopia (pathological)
N-glycan levels
Narcolepsy
Natriuretic peptide levels
Nephrolithiasis
Neuranatomic and neurocognitive phenotypes
Neuroblastoma
Neuroblastoma (high-risk)
Neuroticism
Nevirapine-induced rash
Nevus count
Nicotine dependence

Non-small cell lung cancer
Nonsyndromic cleft lip with or without cleft palate
Normalised brain volume
Obesity
Obesity (early onset extreme)
Obesity (extreme)
Obesity and osteoporosis
Obesity-related traits
Oesophageal cancer
Oesophageal cancer and gastric cancer
Optic disc parameters
Optic disc size (cup)
Optic disc size (disc)
Optic disc size (rim)
Osteoarthritis
Osteonecrosis of the jaw
Osteoporosis
Osteoporosis-related phenotypes
Other erythrocyte phenotypes
Otosclerosis
Ovarian cancer
P-tau181p
Paget's disease
Pain
Pancreatic cancer
Panic disorder
Parkinson's disease
Parkinson's disease (age of onset)
Parkinson's disease (familial)
Parkinson's disease (interaction with coffee consumption)
Partial epilepsies
Peripartum cardiomyopathy
Peripheral artery disease
Permanent tooth development
Personality dimensions
Phospholipid levels (plasma)

Plasma carotenoid and tocopherol levels
Plasma chemerin levels
Plasma coagulation factors
Plasma E-selectin levels
Plasma homocysteine
Plasma level of vitamin B12
Plasma levels of liver enzymes
Plasma levels of polyunsaturated fatty acids
Plasma levels of Protein C
Plasma Lp (a) levels
Plasma vWF and FVIII levels
Platelet aggregation
Platelet counts
Platelet function and related traits
Polycystic ovary syndrome
Postoperative nausea and vomiting
Postoperative ventricular dysfunction
PR interval
Primary tooth development (number of teeth)
Primary tooth development (time to first tooth eruption)
Progranulin levels
Progressive supranuclear palsy
Proinsulin levels
Prostate cancer
Prostate cancer mortality
Protein quantitative trait loci
Pseudoexfoliation syndrome
Pulmonary function
Pulmonary function measures
Pulmonary function traits (other)
QT interval
Quantitative traits
Radiation response
Reasoning
Recombination rate (females)
Recombination rate (males)

Red blood cell traits
Red vs non-red hair colour
Red vs. non-red hair colour
Refractive error
Renal cell carcinoma
Response to acetaminophen (hepatotoxicity)
Response to antidepressant treatment
Response to antidepressants
Response to antineoplastic agents
Response to antipsychotic therapy (extrapyramidal side effects)
Response to antipsychotic treatment
Response to cerivastatin
Response to citalopram treatment
Response to clopidogrel therapy
Response to diuretic therapy
Response to iloperidone treatment (PANSS-T score)
Response to iloperidone treatment (QT prolongation)
Response to lithium treatment in bipolar disorder
Response to metformin
Response to methylphenidate treatment
Response to platinum-based chemotherapy in non-small-cell lung cancer
Response to platinum-based chemotherapy in small cell and non-small cell lung cancers
Response to platinum-based chemotherapy in small-cell lung cancer
Response to statin therapy
Response to ximelagatran treatment
Resting heart rate
Restless legs syndrome
Retinal vascular caliber
Retinol levels
Ribavirin-induced anemia
RR interval (heart rate)
Schizophrenia
Schizophrenia, bipolar disorder and depression (combined)
Scoliosis
Select biomarker traits

Self-rated health
Serum bilirubin levels
Serum butyrylcholinesterase
Serum calcium
Serum creatinine
Serum dehydroepiandrosterone sulphate levels
Serum hepcidin
Serum iron levels
Serum magnesium levels
Serum markers of iron status
Serum metabolites
Serum phosphorus levels
Serum phytosterol levels
Serum prostate-specific antigen levels
Serum urate
Serum uric acid
Sick sinus syndrome
Sickle cell anaemia (severity)
Skin pigmentation
Skin sensitivity to sun
Sleep duration
Sleepiness
Small-cell lung cancer
Smoking behaviour
Smoking cessation
Social and Non-Social Autistic-Like Traits
Speech perception in dyslexia
Sphingolipid levels
Spine bone size
Stroke
Subarachnoid aneurysmal haemorrhage
Subclinical atherosclerosis traits (other)
Subclinical brain infarct
Sudden cardiac arrest
Suicidal ideation
Suicide attempts in bipolar disorder

Systolic blood pressure
Tanning
Tardive dyskinesia
Telomere length
Testicular cancer
Testicular germ cell cancer
Testicular germ cell tumour
Thoracic aortic aneurysms and dissections
Thyroid cancer
Thyroid function
Thyroid stimulating hormone
Thyroid volume
Tonometry
Total ventricular volume
TP53 carriage
Triglycerides
Triglycerides-Blood Pressure (TG-BP)
Two-hour glucose challenge
Type 2 diabetes
Type 2 diabetes and 6 quantitative traits
Type 2 diabetes and other traits
Upper aerodigestive tract cancers
Urinary albumin excretion
Urinary bladder cancer
Urinary metabolites
Uterine fibroids
Vascular endothelial growth factor levels
Venous thromboembolism
Ventricular conduction
Ventricular fibrillation
Vertical cup-disc ratio
Vitamin D insufficiency
Vitamin D levels
Vitamin E levels
Volumetric brain MRI
Waist circumference

Waist Circumference - Triglycerides (WC-TG)
Waist circumference and related phenotypes
Waist-hip ratio
Warfarin maintenance dose
Weight
White matter hypersensitivity burden
Working memory
Wrist bone mass
YKL-40 levels

7.3 Hit-SNPs in T-bet binding Sites

Table 7.1: SNPs in binding sites for the transcription factor T-bet. SNPs that had previously been associated with various traits (trait-associated SNPs) were downloaded from the NHGRI GWAS catalogue and all SNPs in strong LD with these trait-associated SNPs were found. Any SNP that was in a binding site for T-bet was recorded in conjunction with the trait-associated SNP with which it was was in high LD and the trait with which the trait-associated SNP was associated.

rsID of SNP in T-bet Binding Site (Hit-SNP)	rsID of SNP from GWAS (Trait-Associated SNP)	r^2	GWAS Trait
rs1006353	rs4771122	0.810	Body mass index
rs10152590	rs10152591	1.000	Height
rs10152591	rs10152591	1.000	Height
rs10760294	rs10818854	0.861	Polycystic ovary syndrome
rs10808568	rs9792269	1.000	Coeliac disease
rs10850435	rs2338104	0.952	HDL cholesterol
rs10861892	rs8179116	1.000	Conduct disorder (symptom count)
rs10929322	rs10210302	0.975	Crohn's disease
rs10929322	rs2241880	0.951	Crohn's disease
rs10929322	rs3792109	0.938	Crohn's disease
rs10929322	rs3828309	0.963	Crohn's disease
rs10930310	rs6749447	0.967	Blood pressure
rs10995195	rs10995190	0.953	Mammographic density and breast cancer
rs11009175	rs11009175	1.000	Depression–quantitative trait
rs11082995	rs9635963	0.813	Protein quantitative trait loci

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs11130317	rs1042779	0.804	Bipolar disorder
rs11130317	rs2251219	0.986	Major mood disorders
rs11135484	rs2549794	0.832	Crohn's disease
rs11571293	rs3087243	0.940	Type 1 diabetes autoantibodies, Rheumatoid arthritis, Type 1 diabetes
rs12136874	rs4140564	1.000	Knee osteoarthritis
rs12574073	rs1128334	1.000	Systemic lupus erythematosus
rs12574073	rs6590330	1.000	Systemic lupus erythematosus
rs12653750	rs2040704	1.000	Serum IgE levels
rs12653750	rs2244012	1.000	Asthma
rs12928552	rs16965039	1.000	Coronary heart disease
rs12946510	rs11078927	0.852	Asthma
rs12946510	rs2305480	0.852	Asthma, Ulcerative colitis
rs12946510	rs2872507	0.873	Type 1 diabetes autoantibodies, Ulcerative colitis, Rheumatoid arthritis, Crohn's disease
rs12946510	rs907092	0.929	Primary biliary cirrhosis
rs13333528	rs9929218	0.898	Colorectal cancer
rs1335512	rs4636294	1.000	Cutaneous nevi
rs1364340	rs1424233	0.975	Obesity
rs1374910	rs1374910	1.000	Type 2 diabetes
rs1420106	rs13015714	1.000	Coeliac disease
rs1420106	rs2058660	0.980	Crohn's disease
rs1420106	rs917997	1.000	Coeliac disease
rs1465321	rs13015714	1.000	Coeliac disease
rs1465321	rs2058660	1.000	Crohn's disease
rs1465321	rs917997	1.000	Coeliac disease
rs149299	rs151181	0.974	Crohn's disease
rs149299	rs4788084	0.961	Type 1 diabetes autoantibodies, Type I diabetes
rs1551398	rs1551398	1.000	Crohn's disease
rs1551399	rs1551398	0.974	Crohn's disease
rs1610588	rs3129055	0.918	Nasopharyngeal carcinoma
rs1631457	rs1322512	1.000	Tonometry
rs17032405	rs2292303	1.000	Height
rs17032405	rs5742692	1.000	Height
rs17103930	rs8179116	1.000	Conduct disorder (symptom count)
rs17234657	rs17234657	1.000	Crohn's disease
rs17234657	rs4613763	1.000	Crohn's disease
rs1738074	rs1738074	1.000	Multiple sclerosis, Coeliac disease
rs17772411	rs9929218	0.906	Colorectal cancer

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs1788183	rs1371867	0.926	Atrioventricular conduction
rs1806689	rs10785581	0.875	Hypertension
rs1806689	rs7960483	0.847	Hypertension
rs1918788	rs12185268	0.959	Parkinson's disease
rs1918788	rs183211	0.921	Parkinson's disease
rs1918788	rs199533	1.000	Parkinson's disease
rs1918788	rs2942168	0.962	Parkinson's disease
rs1918788	rs393152	0.959	Parkinson's disease
rs1918788	rs415430	0.958	Parkinson's disease
rs1918788	rs8070723	0.959	Progressive supranuclear palsy, Parkinson's disease
rs2070615	rs2070615	1.000	Bipolar disorder
rs2106346	rs1076160	0.988	Psoriasis
rs216172	rs1231206	1.000	Coronary heart disease
rs216172	rs216172	1.000	Coronary heart disease
rs2179225	rs4140564	0.872	Knee Osteoarthritis
rs2268080	rs4911414	0.970	Burning and freckling, Freckles, Red Vs. non-red hair colour, Skin sensitivity to sun
rs2273017	rs2273017	1.000	Graves' disease
rs228599	rs11212617	1.000	Response to metformin
rs2288344	rs8032158	0.922	Keloid
rs2331903	rs1411478	1.000	Progressive supranuclear palsy
rs2387397	rs4750316	0.832	Rheumatoid arthritis
rs2421016	rs6585827	1.000	Height
rs2532234	rs12185268	0.982	Parkinson's disease
rs2532234	rs199533	0.910	Parkinson's disease
rs2532234	rs2074404	0.831	Coeliac disease
rs2532234	rs2942168	1.000	Parkinson's disease
rs2532234	rs393152	1.000	Parkinson's disease
rs2532234	rs415430	0.893	Parkinson's disease
rs2532234	rs8070723	1.000	Progressive supranuclear palsy, Parkinson's disease
rs264834	rs169082	0.810	Protein quantitative trait loci
rs2703078	rs2762051	1.000	Coeliac disease
rs2706383	rs1016988	0.850	Fibrinogen
rs2706383	rs2522056	1.000	Fibrinogen
rs2720665	rs2608053	0.923	Hodgkin's lymphoma
rs2762060	rs806321	0.952	Multiple sclerosis
rs28365932	rs8179116	1.000	Conduct disorder (symptom count)
rs2897908	rs10785581	1.000	Hypertension

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs2897908	rs7960483	1.000	Hypertension
rs2904259	rs7671167	1.000	Chronic obstructive pulmonary disease
rs2984920	rs2816316	0.964	Coeliac disease
rs3024505	rs3024493	1.000	Ulcerative colitis
rs3024505	rs3024505	1.000	Type 1 diabetes autoantibodies, Ulcerative colitis, Crohn's disease, Type I diabetes
rs307896	rs307896	1.000	Multiple sclerosis
rs3091310	rs13098911	1.000	Coeliac disease
rs3131096	rs4324798	0.941	Lung adenocarcinoma
rs3181080	rs13098911	1.000	Coeliac disease
rs354042	rs354033	1.000	Multiple sclerosis
rs36594	rs36600	1.000	Lung cancer
rs3733041	rs2251219	0.922	Major mood disorders
rs3752948	rs3814219	0.985	Endothelial function traits
rs3758354	rs3758354	1.000	Schizophrenia, Bipolar disorder and depression (combined)
rs3784388	rs7178909	0.922	Common traits (other)
rs3789053	rs16832015	1.000	Cognitive performance
rs3810610	rs2283792	1.000	Multiple sclerosis
rs3821383	rs10936599	0.939	Multiple sclerosis, Coeliac disease, Colorectal cancer
rs3821383	rs12696304	1.000	Telomere length
rs4024110	rs3194051	0.870	Ulcerative colitis
rs4265380	rs10903122	0.951	Coeliac disease
rs4265380	rs11249215	0.928	Ankylosing spondylitis
rs4398410	rs10937405	0.886	Lung adenocarcinoma
rs4432068	rs2669010	1.000	Systemic lupus erythematosus
rs4593512	rs6468544	0.971	Antipsychotic-induced QTC interval prolongation
rs4668070	rs13393173	0.831	Response to TNF antagonist treatment
rs4683205	rs9990343	0.804	Brain structure
rs4713186	rs3129109	0.855	Height
rs4713186	rs4947339	0.940	Platelet aggregation
rs4847378	rs12745968	1.000	Bipolar disorder and schizophrenia
rs485499	rs485499	1.000	Primary biliary cirrhosis
rs485789	rs485499	1.000	Primary biliary cirrhosis
rs4906172	rs4906172	1.000	Menopause (age at onset)
rs4944195	rs10899489	0.859	Menarche (age at onset)
rs4944195	rs2373115	0.887	Alzheimer's disease (late onset)
rs5778	rs7832552	0.817	Body mass (lean)

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs631106	rs10889353	0.947	Total cholesterol, LDL cholesterol, Triglycerides
rs631106	rs1167998	0.948	Triglycerides
rs631106	rs1168013	0.974	Triglycerides
rs631106	rs1748195	0.974	Triglycerides
rs6556405	rs1473247	0.982	Mean platelet volume
rs6723449	rs11676348	0.831	Ulcerative colitis
rs6784841	rs6806528	0.975	Coeliac disease
rs7156191	rs10133111	0.910	Brain imaging in schizophrenia (interaction)
rs727263	rs7335046	0.855	Basal cell carcinoma
rs743776	rs743777	1.000	Type 1 diabetes autoantibodies, Rheumatoid arthritis
rs743777	rs743777	1.000	Type 1 diabetes autoantibodies, Rheumatoid arthritis
rs7441808	rs10517086	0.947	Type 1 diabetes
rs7441808	rs874040	0.947	Rheumatoid arthritis
rs7787612	rs6943555	0.831	Alcohol consumption
rs7801282	rs875971	0.987	Aortic root size
rs7850707	rs10816533	1.000	Height
rs7859727	rs10757278	0.877	Myocardial infarction
rs7859727	rs1333048	0.951	Coronary heart disease
rs7859727	rs1333049	0.901	Coronary heart disease
rs7859727	rs2383207	0.905	Abdominal aortic aneurysm
rs7859727	rs4977574	1.000	Coronary heart disease, Myocardial infarction (early onset)
rs7859727	rs944797	0.905	Coronary heart disease
rs7923800	rs11013962	1.000	Common traits (other)
rs8008961	rs911263	1.000	Primary biliary cirrhosis
rs8062727	rs9302752	0.932	Leprosy
rs8129743	rs2014300	0.856	Oesophageal cancer
rs907613	rs907611	0.986	Ulcerative colitis
rs9292777	rs11742570	0.933	Crohn's disease
rs9292777	rs1373692	0.973	Crohn's disease
rs9292777	rs1992660	0.923	Crohn's disease
rs9292777	rs6451493	0.933	Ulcerative colitis
rs9292777	rs6896969	0.949	Multiple sclerosis
rs9292777	rs9292777	1.000	Crohn's disease
rs9300536	rs7335046	0.844	Basal cell carcinoma
rs9303029	rs9303029	1.000	Protein quantitative trait loci

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs9393984	rs2523822	0.972	Drug-induced liver injury (Amoxicillin-Clavulanate)
rs9393984	rs6904029	1.000	Vitiligo
rs9594738	rs9533090	1.000	Bone mineral density (spine)
rs9594738	rs9594738	1.000	Bone mineral density, Bone mineral density (hip)
rs9909593	rs11078927	0.941	Asthma
rs9909593	rs2290400	0.807	Type 1 diabetes
rs9909593	rs2305480	0.941	Asthma, Ulcerative colitis
rs9909593	rs2872507	0.964	Type 1 diabetes autoantibodies, Ulcerative colitis, Crohn's disease, Rheumatoid arthritis
rs9909593	rs7216389	0.803	Asthma
rs9909593	rs8067378	0.860	Ulcerative colitis
rs9909593	rs907092	0.976	Primary biliary cirrhosis
rs9909593	rs9303277	0.886	Primary biliary cirrhosis
rs9944207	rs4886707	1.000	Height

7.4 Hit-SNPs in GATA3 Binding Sites Th1 cells

Table 7.2: SNPs in binding sites for the transcription factor GATA3 in Th1 cells. SNPs that had previously been associated with various traits (trait-associated SNPs) were downloaded from the NHGRI GWAS catalogue and all SNPs in strong LD with these trait-associated SNPs were found. Any SNP that was in a binding site for GATA3 was recorded in conjunction with the trait-associated SNP with which it was was in high LD and the trait with which the trait-associated SNP was associated.

rsID of SNP in GATA3 Binding Site (Hit-SNP)	rsID of GWAS SNP (Trait-Associated SNP)	r^2	GWAS Trait
rs1006353	rs4771122	0.810	Body mass index
rs10152590	rs10152591	1.000	Height
rs10152591	rs10152591	1.000	Height
rs10263639	rs10263639	1.000	Breast cancer
rs10267797	rs886774	0.887	Ulcerative colitis
rs10421601	rs10403021	1.000	Diabetic retinopathy
rs10828247	rs11012732	0.828	Meningioma
rs10861892	rs8179116	1.000	Conduct disorder (symptom count)
rs10929322	rs10210302	0.975	Crohn's disease
rs10929322	rs2241880	0.951	Crohn's disease
rs10929322	rs3792109	0.938	Crohn's disease
rs10929322	rs3828309	0.963	Crohn's disease
rs10930310	rs6749447	0.967	Blood pressure
rs10995195	rs10995190	0.953	Mammographic density, Breast cancer
rs11130317	rs1042779	0.804	Bipolar disorder
rs11130317	rs2251219	0.986	Major mood disorders
rs11135484	rs2549794	0.832	Crohn's disease
rs11142	rs1933182	0.915	Chronic kidney disease
rs11742570	rs11742570	1.000	Crohn's disease
rs11742570	rs1373692	0.960	Crohn's disease
rs11742570	rs1992660	1.000	Crohn's disease
rs11742570	rs6451493	1.000	Ulcerative colitis
rs11742570	rs6896969	0.974	Multiple sclerosis
rs11742570	rs9292777	0.933	Crohn's disease
rs12038333	rs1329424	1.000	Age-related macular degeneration
rs12132349	rs11584383	0.855	Ulcerative colitis, Crohn's disease
rs12132349	rs2297909	0.881	Ankylosing spondylitis
rs12132349	rs296547	0.803	Coeliac disease
rs12132349	rs7554511	0.971	Ulcerative colitis
rs1279750	rs1279750	1.000	Platelet counts

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs12991737	rs3771180	0.806	Asthma
rs1323292	rs2816316	1.000	Coeliac disease
rs133015	rs738322	0.872	Cutaneous nevi
rs133016	rs738322	1.000	Cutaneous nevi
rs133017	rs738322	1.000	Cutaneous nevi
rs1335512	rs4636294	1.000	Cutaneous nevi
rs13418039	rs1559040	1.000	Sudden cardiac arrest
rs1364340	rs1424233	0.975	Obesity
rs1406069	rs12185268	1.000	Parkinson's disease
rs1406069	rs183211	0.914	Parkinson's disease
rs1406069	rs199533	0.957	Parkinson's disease
rs1406069	rs2942168	0.960	Parkinson's disease
rs1406069	rs393152	1.000	Parkinson's disease
rs1406069	rs415430	0.914	Parkinson's disease
rs1406069	rs8070723	1.000	Progressive supranuclear palsy, Parkinson's disease
rs1420106	rs13015714	1.000	Coeliac disease
rs1420106	rs2058660	0.980	Crohn's disease
rs1420106	rs917997	1.000	Coeliac disease
rs1445899	rs1445898	0.987	Type 1 diabetes
rs1465321	rs13015714	1.000	Coeliac disease
rs1465321	rs2058660	1.000	Crohn's disease
rs1465321	rs917997	1.000	Coeliac disease
rs1481892	rs900145	0.937	Menarche (age at onset)
rs1551398	rs1551398	1.000	Crohn's disease
rs1551399	rs1551398	0.974	Crohn's disease
rs1570069	rs1321535	0.882	Phospholipid levels (plasma)
rs1570069	rs2236212	0.941	Phospholipid levels (plasma)
rs1570069	rs3734398	0.953	Phospholipid levels (plasma)
rs1570069	rs3798713	0.952	Phospholipid levels (plasma)
rs1570069	rs4711171	0.893	Phospholipid levels (plasma)
rs1570069	rs4713103	0.830	Phospholipid levels (plasma)
rs1570069	rs6918936	0.828	Phospholipid levels (plasma)
rs1596017	rs10210302	0.951	Crohn's disease
rs1596017	rs2241880	0.904	Crohn's disease
rs1596017	rs3792109	0.877	Crohn's disease
rs1596017	rs3828309	0.928	Crohn's disease
rs1610588	rs3129055	0.918	Nasopharyngeal carcinoma
rs17035375	rs17035378	0.985	Coeliac disease
rs17035378	rs17035378	1.000	Coeliac disease

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs17103930	rs8179116	1.000	Conduct disorder (symptom count)
rs17411033	rs4598195	0.947	Ulcerative colitis
rs17533167	rs10484561	0.894	Follicular lymphoma
rs1859156	rs1859156	1.000	Attention deficit hyperactivity disorder
rs1886730	rs734999	0.906	Ulcerative colitis
rs2016755	rs738322	0.875	Cutaneous nevi
rs2058622	rs13015714	1.000	Coeliac disease
rs2058622	rs2058660	1.000	Crohn's disease
rs2058622	rs917997	1.000	Coeliac disease
rs2106346	rs1076160	0.988	Psoriasis
rs2145998	rs2145998	1.000	Height
rs216172	rs1231206	1.000	Coronary heart disease
rs216172	rs216172	1.000	Coronary heart disease
rs2267020	rs5751614	1.000	Height
rs2275271	rs7914558	1.000	Schizophrenia
rs2278563	rs4907240	0.936	Event-related brain oscillations
rs2282030	rs8017423	0.927	Mortality among heart failure patients
rs231727	rs1024161	0.829	Graves' disease, Alopecia areata
rs2532234	rs12185268	0.982	Parkinson's disease
rs2532234	rs199533	0.910	Parkinson's disease
rs2532234	rs2074404	0.831	Coeliac disease
rs2532234	rs2942168	1.000	Parkinson's disease
rs2532234	rs393152	1.000	Parkinson's disease
rs2532234	rs415430	0.893	Parkinson's disease
rs2532234	rs8070723	1.000	Progressive supranuclear palsy, Parkinson's disease
rs2532296	rs12185268	1.000	Parkinson's disease
rs2532296	rs183211	0.882	Parkinson's disease
rs2532296	rs199533	0.959	Parkinson's disease
rs2532296	rs2942168	1.000	Parkinson's disease
rs2532296	rs393152	1.000	Parkinson's disease
rs2532296	rs415430	0.919	Parkinson's disease
rs2532296	rs8070723	1.000	Progressive supranuclear palsy, Parkinson's disease
rs261360	rs261360	1.000	Hair morphology
rs2639186	rs2567426	0.800	Information processing speed
rs2702136	rs6707600	1.000	Working memory
rs2703078	rs2762051	1.000	Coeliac disease
rs2706383	rs1016988	0.850	Fibrinogen
rs2706383	rs2522056	1.000	Fibrinogen

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs2762060	rs806321	0.952	Multiple sclerosis
rs281438	rs281437	1.000	Soluble ICAM-1
rs2865530	rs2865531	1.000	Pulmonary function
rs286913	rs286913	1.000	Response to antipsychotic treatment
rs3101018	rs3117582	0.962	Lung adenocarcinoma, Lung cancer
rs3101018	rs3131379	0.962	Systemic lupus erythematosus
rs3129763	rs3129763	1.000	Systemic sclerosis
rs3129763	rs602875	1.000	Leprosy
rs3129763	rs674313	0.904	Chronic lymphocytic leukaemia
rs3131096	rs4324798	0.941	Lung adenocarcinoma
rs3213621	rs11130248	1.000	Keloid
rs3732123	rs17027258	1.000	White blood cell types
rs3738398	rs12042938	0.809	Neuranatomic and neurocognitive phenotypes
rs3758253	rs755109	1.000	Quantitative traits
rs3758354	rs3758354	1.000	Schizophrenia, bipolar disorder and depression (combined)
rs3788013	rs3788013	1.000	Type 1 diabetes autoantibodies
rs3788013	rs9976767	0.865	Type 1 diabetes
rs3806288	rs6702784	0.852	Diabetic retinopathy
rs3863439	rs10871290	0.814	Breast cancer
rs40452	rs42490	1.000	Leprosy
rs4432068	rs2669010	1.000	Systemic lupus erythematosus
rs4522587	rs4675374	1.000	Coeliac disease
rs460901	rs410644	0.813	Anorexia nervosa
rs4667114	rs1018326	0.927	Ankylosing spondylitis
rs4667114	rs13010713	0.882	Coeliac disease
rs4668070	rs13393173	0.831	Response to TNF antagonist treatment
rs4679048	rs2070488	1.000	Electrocardiographic conduction measures
rs4683205	rs9990343	0.804	Brain structure
rs4686760	rs4686760	1.000	Plasma vWF and FVIII levels
rs4713581	rs6457617	0.952	Graves' disease, Systemic sclerosis, Rheumatoid arthritis
rs4713581	rs6457620	0.952	Rheumatoid arthritis
rs4713582	rs6457617	1.000	Graves' disease, Systemic sclerosis, Rheumatoid arthritis
rs4713582	rs6457620	1.000	Rheumatoid arthritis
rs4772190	rs7335046	0.855	Basal cell carcinoma
rs4888376	rs2865531	0.978	Pulmonary function
rs4917014	rs4917014	1.000	Systemic lupus erythematosus

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs507101	rs7928794	1.000	Response to antipsychotic therapy (extrapyramidal side effects)
rs6441957	rs9990343	0.925	Brain structure
rs6651252	rs10088218	1.000	Ovarian cancer
rs6651252	rs6651252	1.000	Crohn's disease
rs6686643	rs6686643	1.000	Total ventricular volume
rs6919110	rs12199222	0.865	Height
rs6997	rs3197999	0.867	Primary sclerosing cholangitis, Crohn's disease, Ulcerative colitis
rs6997	rs9822268	0.897	Ulcerative colitis
rs6997	rs9858542	0.897	Ulcerative colitis, Crohn's Disease
rs7004723	rs6984045	1.000	Multiple sclerosis
rs707952	rs3129763	0.815	Systemic sclerosis
rs707952	rs602875	0.815	Leprosy
rs707952	rs674313	0.815	Chronic lymphocytic leukaemia
rs707952	rs9272219	0.836	Rheumatoid arthritis, Schizophrenia
rs707952	rs9272535	0.872	Chronic lymphocytic leukaemia
rs7090445	rs10821936	0.972	Acute lymphoblastic leukaemia (childhood)
rs7090445	rs7089424	0.902	Acute lymphoblastic leukaemia (childhood)
rs7117353	rs10437629	0.852	Suicide attempts in bipolar disorder
rs7283760	rs4819388	0.849	Coeliac disease
rs734252	rs2677744	0.920	Attention deficit hyperactivity disorder
rs738322	rs738322	1.000	Cutaneous nevi
rs7578035	rs7578035	1.000	Bipolar disorder
rs7585265	rs2286963	0.842	Serum metabolites
rs7734434	rs17234657	1.000	Crohn's disease
rs7734434	rs4613763	1.000	Crohn's disease
rs7965349	rs7957197	0.835	Type 2 diabetes
rs8014856	rs2119704	1.000	Multiple sclerosis
rs8015102	rs2119704	1.000	Multiple sclerosis
rs8027781	rs6494537	0.962	Haematological and biochemical traits
rs8192917	rs8192917	1.000	Vitiligo
rs851984	rs2941740	0.821	Bone mineral density (hip)
rs859648	rs859637	0.976	Coeliac disease
rs881375	rs1953126	1.000	Coeliac disease and Rheumatoid arthritis
rs881375	rs881375	1.000	Rheumatoid arthritis
rs907613	rs907611	0.986	Ulcerative colitis
rs9271170	rs9271100	1.000	Systemic lupus erythematosus
rs9271612	rs3129763	0.948	Systemic sclerosis
rs9271612	rs602875	0.948	Leprosy

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs9271612	rs674313	0.948	Chronic lymphocytic leukaemia
rs9271613	rs3129763	0.896	Systemic sclerosis
rs9271613	rs602875	0.896	Leprosy
rs9271613	rs674313	0.895	Chronic lymphocytic leukaemia
rs9303029	rs9303029	1.000	Protein quantitative trait loci
rs9314099	rs17376456	0.949	Diabetic retinopathy
rs9314100	rs17376456	0.948	Diabetic retinopathy
rs9393984	rs2523822	0.972	Drug-induced liver injury (amoxicillin-clavulanate)
rs9393984	rs6904029	1.000	Vitiligo
rs9442234	rs9442235	0.975	Cognitive performance
rs9909593	rs11078927	0.941	Asthma
rs9909593	rs2290400	0.807	Type 1 diabetes
rs9909593	rs2305480	0.941	Asthma, Ulcerative colitis
rs9909593	rs2872507	0.964	Type 1 diabetes autoantibodies, Ulcerative colitis, Rheumatoid arthritis, Crohn's disease
rs9909593	rs7216389	0.803	Asthma
rs9909593	rs8067378	0.860	Ulcerative colitis
rs9909593	rs907092	0.976	Primary biliary cirrhosis
rs9909593	rs9303277	0.886	Primary biliary cirrhosis

7.5 Hit-SNPs in GATA3 Binding Sites in Th2 cells

Table 7.3: SNPs in binding sites for the transcription factor GATA3 in Th2 cells. SNPs that had previously been associated with various traits (trait-associated SNPs) were downloaded from the NHGRI GWAS catalogue and all SNPs in strong LD with these trait-associated SNPs were found. Any SNP that was in a binding site for GATA3 was recorded in conjunction with the trait-linked SNP with which it was was in high LD and the trait with which the trait-associated SNP was associated.

rsID of SNP in GATA3 Binding Site (Hit-SNP)	rsID of GWAS SNP (Trait-Associated SNP)	r^2	GWAS Trait
rs1006353	rs4771122	0.810	Body mass index
rs10152591	rs10152591	1.000	Height
rs10267797	rs886774	0.887	Ulcerative colitis
rs10516286	rs16891867	1.000	Conduct disorder (symptom count)
rs10516286	rs1861046	1.000	Conduct disorder (case status)
rs10516286	rs1861050	1.000	Conduct disorder (case status)
rs10899476	rs10899489	0.927	Menarche (age at onset)
rs10899476	rs2373115	1.000	Alzheimer's disease (late onset)
rs10929322	rs10210302	0.975	Crohn's disease
rs10929322	rs2241880	0.951	Crohn's disease
rs10929322	rs3792109	0.938	Crohn's disease
rs10929322	rs3828309	0.963	Crohn's disease
rs10930310	rs6749447	0.967	Blood pressure
rs10995195	rs10995190	0.953	Mammographic density, Breast cancer
rs11130317	rs1042779	0.804	Bipolar disorder
rs11130317	rs2251219	0.986	Major mood disorders
rs11135484	rs2549794	0.832	Crohn's disease
rs11142	rs1933182	0.915	Chronic kidney disease
rs11663558	rs1805081	0.834	Obesity
rs11692725	rs882300	0.976	Multiple sclerosis, Electrocardiographic traits
rs12132349	rs11584383	0.855	Ulcerative colitis, Crohn's disease
rs12132349	rs2297909	0.881	Ankylosing spondylitis
rs12132349	rs296547	0.803	Coeliac disease
rs12132349	rs7554511	0.971	Ulcerative colitis
rs12252379	rs7913069	1.000	Uterine fibroids
rs1279750	rs1279750	1.000	Platelet counts
rs1285403	rs1285407	1.000	Protein quantitative trait loci
rs12942708	rs12936587	0.841	Coronary heart disease
rs12991737	rs3771180	0.806	Asthma
rs1323292	rs2816316	1.000	Coeliac disease

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs1327235	rs1327235	1.000	Diastolic blood pressure, Systolic blood pressure, Blood pressure
rs1388387	rs7578597	1.000	Type 2 diabetes
rs1406069	rs12185268	1.000	Parkinson's disease
rs1406069	rs183211	0.914	Parkinson's disease
rs1406069	rs199533	0.957	Parkinson's disease
rs1406069	rs2942168	0.960	Parkinson's disease
rs1406069	rs393152	1.000	Parkinson's disease
rs1406069	rs415430	0.914	Parkinson's disease
rs1406069	rs8070723	1.000	Progressive supranuclear palsy, Parkinson's disease
rs1420106	rs13015714	1.000	Coeliac disease
rs1420106	rs2058660	0.980	Crohn's disease
rs1420106	rs917997	1.000	Coeliac disease
rs1596017	rs10210302	0.951	Crohn's disease
rs1596017	rs2241880	0.904	Crohn's disease
rs1596017	rs3792109	0.877	Crohn's disease
rs1596017	rs3828309	0.928	Crohn's disease
rs1610588	rs3129055	0.918	Nasopharyngeal carcinoma
rs17103930	rs8179116	1.000	Conduct disorder (symptom count)
rs17753121	rs7336332	0.961	Weight
rs1859156	rs1859156	1.000	Attention deficit hyperactivity disorder
rs1886730	rs734999	0.906	Ulcerative colitis
rs1893447	rs10899489	0.859	Menarche (age at onset)
rs1893447	rs2373115	0.887	Alzheimer's disease (late onset)
rs1983351	rs7495052	1.000	Inattentive symptoms
rs2045592	rs442177	0.825	Triglycerides
rs2058622	rs13015714	1.000	Coeliac disease
rs2058622	rs2058660	1.000	Crohn's disease
rs2058622	rs917997	1.000	Coeliac disease
rs2106346	rs1076160	0.988	Psoriasis
rs2145998	rs2145998	1.000	Height
rs2239826	rs3093023	0.803	Rheumatoid arthritis
rs2239826	rs3093024	0.827	Rheumatoid arthritis
rs2275271	rs7914558	1.000	Schizophrenia
rs2278563	rs4907240	0.936	Event-related brain oscillations
rs2287900	rs3194051	0.918	Ulcerative colitis
rs231727	rs1024161	0.829	Graves' disease, Alopecia areata
rs2408025	rs11958779	0.823	Height
rs281438	rs281437	1.000	Soluble ICAM-1

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs286913	rs286913	1.000	Response to antipsychotic treatment
rs2984920	rs2816316	0.964	Coeliac disease
rs3129763	rs3129763	1.000	Systemic sclerosis
rs3129763	rs602875	1.000	Leprosy
rs3129763	rs674313	0.904	Chronic lymphocytic leukaemia
rs34251816	rs10026364	1.000	Coronary heart disease
rs34361	rs2112347	0.848	Body mass index
rs3732123	rs17027258	1.000	White blood cell types
rs3758253	rs755109	1.000	Quantitative traits
rs3758354	rs3758354	1.000	Schizophrenia, bipolar disorder and depression (combined)
rs3792112	rs10210302	0.975	Crohn's disease
rs3792112	rs2241880	0.975	Crohn's disease
rs3792112	rs3792109	0.952	Crohn's disease
rs3792112	rs3828309	1.000	Crohn's disease
rs4024110	rs3194051	0.870	Ulcerative colitis
rs40452	rs42490	1.000	Leprosy
rs41156	rs5753037	0.975	Type 1 diabetes
rs4522587	rs4675374	1.000	Coeliac disease
rs460901	rs410644	0.813	Anorexia nervosa
rs4668070	rs13393173	0.831	Response to TNF antagonist treatment
rs4679048	rs2070488	1.000	Electrocardiographic conduction measures
rs4683205	rs9990343	0.804	Brain structure
rs4772190	rs7335046	0.855	Basal cell carcinoma
rs4939490	rs4939490	1.000	Multiple sclerosis
rs507101	rs7928794	1.000	Response to antipsychotic therapy (extrapyramidal side effects)
rs6441957	rs9990343	0.925	Brain structure
rs6478486	rs1953126	1.000	Coeliac disease and Rheumatoid arthritis
rs6478486	rs881375	1.000	Rheumatoid arthritis
rs6556405	rs1473247	0.982	Mean platelet volume
rs6604026	rs6604026	1.000	Multiple sclerosis
rs6685472	rs12733856	0.885	Mortality among heart failure patients
rs6936240	rs11968814	1.000	F-cell distribution
rs7004723	rs6984045	1.000	Multiple sclerosis
rs7165988	rs12899449	0.957	Bipolar disorder
rs7165988	rs12912251	1.000	Bipolar disorder, Bipolar disorder and major depressive disorder (combined)
rs7171233	rs12899449	0.957	Bipolar disorder

Hit-SNP	Trait-Associated SNP	r^2	Trait
rs7171233	rs12912251	1.000	Bipolar disorder, Bipolar disorder and major depressive disorder (combined)
rs727263	rs7335046	0.855	Basal cell carcinoma
rs7488228	rs10858945	0.976	Optic disc size (cup)
rs7578035	rs7578035	1.000	Bipolar disorder
rs7585265	rs2286963	0.842	Serum metabolites
rs7602799	rs7602460	1.000	Atrioventricular conduction
rs7705826	rs6596075	0.942	Crohn's disease
rs7988412	rs4771122	0.858	Body mass index
rs8027781	rs6494537	0.962	Haematological and biochemical traits
rs8062727	rs9302752	0.932	Leprosy
rs809930	rs3774372	0.823	Blood pressure
rs851984	rs2941740	0.821	Bone mineral density (hip)
rs881375	rs1953126	1.000	Coeliac disease and Rheumatoid arthritis
rs881375	rs881375	1.000	Rheumatoid arthritis
rs907613	rs907611	0.986	Ulcerative colitis
rs9271612	rs3129763	0.948	Systemic sclerosis
rs9271612	rs602875	0.948	Leprosy
rs9271612	rs674313	0.948	Chronic lymphocytic leukaemia
rs9271613	rs3129763	0.896	Systemic sclerosis
rs9271613	rs602875	0.896	Leprosy
rs9271613	rs674313	0.895	Chronic lymphocytic leukaemia
rs933672	rs11958779	0.823	Height
rs9869432	rs7631605	0.875	P-tau181p

7.6 Genes Located Near Hit-SNPs for T-bet

Table 7.4: Genes located in or near SNP-hits for T-bet. We recorded any genes containing a hit-SNP for T-bet in addition to any genes within 2kbp of a hit-SNP and any genes within 50kbp of a hit-SNP.

rsID of Hit-SNP	Genes overlapping Hit-SNP	Genes within 2kbp of Hit-SNP	Genes within 50kbp of Hit-SNP
rs1006353	none	none	MTIF3, GTF3A
rs10152590	none	none	none
rs10152591	none	none	none
rs10760294	DENND1A (intron)	DENND1A	DENND1A
rs10808568	none	none	none
rs10850435	MVK	MVK	MVK, MMAB
rs10861892	CMKLR1 (intron)	CMKLR1	CMKLR1
rs10929322	none	none	ATG16L1, INPP5D, SCARNA5, SCARN6
rs10930310	STK39 (intron)	STK39	STK39
rs10995195	ZNF365 (intron)	ZNF365	ZNF365
rs11009175	none	none	ITGB1
rs11082995	OSBPL1A (intron)	OSBPL1A	OSBPL1A, CABYR, TTC39C
rs11130317	GLT8D1 (intron)	GLT8D1	GLT8D1, GNL3, PBRM1, SPCS1, NEK4
rs11135484	ERAP2 (intron)	ERAP2	ERAP2, LNPEP
rs11571293	none	none	CTLA4
rs12136874	none	none	none
rs12574073	none	none	ETS1
rs12653750	RAD50 (intron)	RAD50	RAD50, IL13, IL4
rs12928552	NLRC5 (intron)	NLRC5	NLRC5
rs12946510	none	none	IKZF3, GRB7, C17orf37, mir-4728, ERBB2
rs13333528	CDH1 (intron)	CDH1	CDH1
rs1335512	none	none	MTAP
rs1364340	none	none	none
rs1374910	IGF2BP2 (intron)	IGF2BP2	IGF2BP2
rs1420106	none	IL18RAP	IL18RAP, IL18R1, mir-4772
rs1465321	IL18R1 (intron)	IL18R1	IL18R1, IL1RL1, IL18RAP
rs149299	none	none	CLN3, APOBR, IL27, EIF3CL
rs1551398	none	none	none
rs1551399	none	none	none
rs1610588	none	none	none
rs1631457	SYNE1 (intron)	SYNE1	SYNE1

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs17032405	none	none	none
rs17103930	none	none	none
rs17234657	none	none	none
rs1738074	TAGAP (5′ UTR)	TAGAP	TAGAP, RSPH3
rs17772411	CDH1 (intron)	CDH1	CDH1
rs1788183	none	none	RNF19A
rs1806689	none	none	none
rs1918788	KIAA1267 (intron)	KIAA1267	KIAA1267
rs2070615	CACNB3 (intron)	CACNB3	CACNB3
rs2106346	TSC1 (intron)	TSC1	TSC1, C9orf9, AK8
rs216172	SMG6 (intron)	SMG6	SMG6
rs2179225	none	none	none
rs2268080	RALV (intron)	RALV	RALV, mir-4755
rs2273017	C6orf10 (intron)	C6orf10	C6orf10, BTNL2
rs228599	ATM (intron)	ATM	NPAT
rs2288344	NEDD4 (intron)	NEDD4	NEDD4
rs2331903	STX6 (intron)	STX6	STX6, KIAA1614
rs2387397	none	none	none
rs2421016	PLEKHA1 (intron)	PLEKHA1	PLEKHA1, mir-3941, ARMS2
rs2532234	KIAA1267 (intron)	KIAA1267	KIAA1267
rs264834	DOCK2 (intron)	DOCK2	DOCK2
rs2703078	none	none	none
rs2706383	C5orf56 (intron)	C5orf56	C5orf56, IRF1
rs2720665	PVT1 oncogene	PVT1 oncogene	PVT1 oncogene, mir-1207
rs2762060	none	none	none
rs28365932	none	DDX54, C12orf52	DDX54, C12orf52, RASAL1, CCDC42B, TPCN1
rs2897908	none	none	none
rs2904259	FAM13A (intron)	FAM13A	FAM13A
rs2984920	none	RGS1	RGS1
rs3024505	none	IL10	IL10, IL19, MAPKAPK2
rs307896	SAE1 (intron)	SAE1	SAE1, ZC3H4
rs3091310	CCR3 (intron)	CCR3	CCR3
rs3131096	TPK1 (intron)	TPK1	TPK1
rs3181080	none	CCR1	CCR1, CCR3
rs354042	none	none	none
rs36594	MTMR3 (intron)	MTMR3	MTMR3
rs3733041	GLT8D1 (intron)	GLT8D1	GLT8D1, GNL3, PBRM1, SPCS1, NEK4
rs3752948	OBFC1 (intron)	OBFC1	OBFC1
rs3758354	none	none	ANXA1

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs3784388	AP3S2 (5′ UTR)	AP3S2	AP3S2, C15orf38
rs3789053	MDM4 (intron)	MDM4	MDM4, PIK3C2B
rs3810610	MAPK1 (3′ UTR)	MAPK1	MPAK1, YPEL1
rs3821383	none	MYNN	MYNN, ARPM1, TERC, LRRC34, LRRIQ4
rs4024110	CAPSL (intron)	CAPSL	CAPSL, UGT3A1, IL7R
rs4265380	none	RUNX3	RUNX3
rs4398410	none	none	TP63
rs4432068	none	none	OSBPL8
rs4593512	none	none	TSPYL5
rs4668070	CERS6 (intron)	CERS6	CERS6, mir-4774
rs4683205	none	none	CCR3
rs4713186	none	C6orf100	C6orf100, TRIM27
rs4847378	FAM69A (intron)	FAM69A	FAM69A, RPL5
rs485499	none	none	IL12A
rs485789	none	none	IL12A
rs4906172	DYNC1H1 (intron)	DYNC1H1	DYNC1H1
rs4944195	GAB2 (intron)	GAB2	GAB2
rs5778	none	TRHR	TRHR
rs631106	none	USP1	USP1, DOCK7
rs6556405	RNF145 (5′UTR)	RNF145	RNF145
rs6723449	CXCR2 (intron)	CXCR2	CXCR2, CXCR1, RUFY4
rs6784841	FRMD4B (intron)	FRMD4B	FRMD4B
rs7156191	TRAF3 (intron)	TRAF3	TRAF3
rs727263	UBAC2 (intron)	UBAC2	UBAC2, mir-623, GPR183
rs743776	none	none	IL2RB, C1QTNF6
rs743777	none	none	IL2RB, C1QTNF6
rs7441808	none	none	none
rs7787612	AUTS2 (intron)	AUTS2	AUTS2
rs7801282	CRCP (intron)	CRCP	CRCP
rs7850707	none	none	ZNF510, ZNF782
rs7859727	CDKN2B antisense RNA	CDKN2B antisense RNA	CDKN2B antisense RNA
rs7923800	KIAA1267 (intron)	KIAA1267	KIAA1267, PRINS
rs8008961	RAD51B (intron)	RAD51B	RAD51B
rs8062727	none	none	NKD1, SNX20, NOD2
rs8129743	Runx1 (intron)	Runx1	Runx1, C21orf96
rs907613	LSP1 (5′ UTR)	LSP1	LSP1, SYT8, TNNI2, mir-4298
rs9292777	none	none	none
rs9300536	UBAC2 (intron)	UBAC2	UBAC2, GPR183, GPR18
rs9303029	none	C17orf62	C17orf62, NARF, C17orf101, HEXDC, LOC100505970

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs9393984	none	none	HCG8, HCG9, HLA-A, HLA-A29.1
rs9594738	none	none	none
rs9909593	IKZF3 (intron)	IKZF3	IKZF3
rs9944207	SNUPN (intron)	SNUPN	SNUPN, SNX33, IMP3, PTPN9

7.7 Genes Located Near Hit-SNPs for GATA3 in Th1 cells

Table 7.5: Genes located in or near SNP-hits for GATA3 in Th1 Cells. We recorded any genes containing a hit-SNP for GATA3 in addition to any genes within 2kbp of a hit-SNP and any genes within 50kbp of a hit-SNP.

rsID of Hit-SNP	Genes overlapping Hit-SNP	Genes with 2kbp of Hit-SNP	Genes within 50kbp of Hit-SNP
rs1006353	none	none	MTIF3, GTF3A
rs10152590	none	none	none
rs10152591	none	none	none
rs10263639	none	none	none
rs10267797	none	none	DLD, LAMB1
rs10421601	none	none	POP4, VSTM2B
rs10828247	none	MLLT10	MLLT10, C10orf140, C10orf114, mir-1915
rs10861892	CMKLR1 (intron)	CMKLR1	CMKLR1
rs10929322	none	none	ATG16L1, INPP5D
rs10930310	STK39 (intron)	STK39	STK39
rs10995195	ZNF365 (intron)	ZNF365	ZNF365
rs11130317	GLT8D1 (intron)	GLT8D1	GLT8D1, GNL3, PBRM1, SPCS1, NEK4
rs11135484	ERAP2 (intron)	ERAP2	ERAP2, LNPEP
rs11142	SORT1 (exon)	SORT1	SORT1, PSMA5, MVBPHL
rs11742570	none	none	none
rs12038333	CFH((intron)	CFH	CFH
rs12132349	C1orf106 (intron)	C1orf106	C1orf106, GPR25, CAMSAP1L1
rs1279750	none	none	TRIM7, OR2V2
rs12991737	none	none	IL18RAP, IL18R1, mir-4772
rs1323292	none	none	RGS1
rs133015	PLA2G6 (intron)	PLA2G6	PLA2G6, MAFF, TMEM184B
rs133016	PLA2G6 (intron)	PLA2G6	PLA2G6, MAFF, TMEM184B
rs133017	PLA2G6 (intron)	PLA2G6	PLA2G6, MAFF, TMEM184B
rs1335512	none	none	MTAP
rs13418039	none	none	ACYP2
rs1364340	none	none	none
rs1406069	KIAA1267 (intron)	KIAA1267	KIAA1267
rs1420106	none	IL18RAP	IL18RAP, IL18R1, mir-4772
rs1445899	CAPSL (intron)	CAPSL	CAPSL, UGT3A1, IL7R
rs1465321	IL18R1 (intron)	IL18R1	IL18R1, IL1RL1, IL18RAP
rs1481892	ARNTL (intron)	ARNTL	ARNTL

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs1551398	none	none	none
rs1551399	none	none	none
rs1570069	ELOVL2 (intron)	ELOVL2	ELOVL2, SYCP2L, LOC100506409
rs1596017	none	none	ATG16L1, INPP5D, SCARNA5, SCARN6
rs1610588	none	none	none
rs17035375	PLEK (intron)	PLEK	PLEK
rs17035378	PLEK (intron)	PLEK	PLEK
rs17103930	none	none	none
rs17411033	none	none	DLD,LAMB1
rs17533167	none	none	HLA-DQB1, HLA-DQA1, HLA-DRB4
rs1859156	BMPR1B (intron)	BMPR1B	BMPR1B
rs1886730	TNFRSF14 (intron)	TNFRSF14	TNFRSF14, MMEL1, C1orf93, HES5, PANK4
rs2016755	PLA2G6 (intron)	PLA2G6	PLA2G6, BAIAP2L2
rs2058622	IL18R1 (intron)	IL18R1	IL18R1, IL1RL1, IL18RAP
rs2106346	TSC1 (intron)	TSC1	TSC1, C9orf9, AK8
rs2145998	none	none	ZMIZ1, PPIF, ZCCHC24
rs216172	SMG6 (intron)	SMG6	SMG6
rs2267020	BCR (intron)	BCR	BCR
rs2275271	CNNM2 (exon)	CNNM2	CNNM2, NT5C2
rs2278563	ARID5A (intron)	ARID5A	ARID5A, KIAA1310
rs2282030	PMSC1 (intron)	PMSC1	PMSC1, C14orf102
rs231727	none	none	CTLA4
rs2532234	KIAA1267 (intron)	KIAA1267	KIAA1267
rs2532296	KIAA1267 (intron)	KIAA1267	KIAA1267
rs261360	none	none	C20orf30
rs2639186	none	none	none
rs2702136	ITSN2 (intron)	ITSN2	ITSN2
rs2703078	none	none	none
rs2706383	C5orf56 (intron)	C5orf56	C5orf56, IRF1
rs2762060	none	none	none
rs281438	none	ICAM4, ICAM5	ICAM4, ICAM5, ICAM1, MRPL4, ZGLP1, FDX1L, RAVER1, ICAM3
rs2865530	CFDP1 (intron)	CFDP1	CFDP1
rs286913	EHF (intron)	EHF	EHF
rs3101018	none	CLIC1, MSH5	CLIC1, MSH5, ABHD16A, LY6G6F, LY6G6D, LY6G6C, C6orf25, DDAH2, VARS, C6orf27, C6orf26

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs3129763	none	none	HLA-DQB1, HLA-DQA1, HLA-DRB4
rs3131096	TPK1 (intron)	TPK1	TPK1
rs3213621	ZNYMD10 (3′)	ZNYMD10, RASSF1	ZNYMD10, RASSF1, IFRD2, HYAL3, NAT6, HYAL1, HYAL2, TUSC2, NPRL2, CYB561D2, TMEM115, CACNA2D2
rs3732123	none	none	IL1RL1, IL18R1, IL18RAP, mir-4772
rs3738398	none	DISC1	DISC1
rs3758253	HEMGN (5′)	HEMGN	HEMGN, ANP32B, C9orf156
rs3758354	none	none	ANXA1
rs3788013	UBASH3A (intron)	UBASH3A	UBASH3A, TMPRSS3
rs3806288	none	MRPS15, CSF3R	MRPS15, CSF3R, OSCP1
rs3863439	none	none	GLG1, CLEC18B
rs40452	RIPK2 (intron)	RIPK2	RIPK2
rs4432068	none	none	OSBPL8
rs4522587	ICOS (intron)	ICOS	ICOS
rs460901	none	none	SSBP2
rs4667114	none	none	none
rs4668070	CERS6 (intron)	CERS6	CERS6, mir-4774
rs4679048	none	none	XYLB
rs4683205	none	none	CCR3
rs4686760	none	none	VPS8
rs4713581	none	none	HLA-DQA2, HLA-DQB1, HLA-DQA1
rs4713582	none	none	HLA-DQA2, HLA-DQB1, HLA-DQA1
rs4772190	UBAC2 (intron)	UBAC2	UBAC2, FKSG29, mir-623
rs4888376	none	none	CFDP1, BCAR1
rs4917014	none	none	IKZF1
rs507101	SLC3A2 (intron)	SLC3A2	SLC3A2, WDR74, RNU2-2, SNHG1, SNORD22, 26, 28, 30, 25, 27, 31 and 25
rs6441957	none	none	CCR3
rs6651252	none	none	none
rs6686643	MGST3 (intron)	MGST3	MGST3, ALDH9A1
rs6919110	FAM8A1 (5′ UTR)	FAM8A1	FAM8A1, NUP153, CAP2
rs6997	TCTA (3′ UTR)	TCTA, AMT	TCTA, AMT, NICN1, RHOA
rs7004723	ASAP1 (intron)	ASAP1	ASAP1
rs707952	HLA-DQA1 (exon)	HLA-DQA1	HLA-DQA1, HLA-DQB1

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs7090445	ARID5B (intron)	ARID5B	ARID5B
rs7117353	none	C11orf41	C11orf41
rs7283760	none	none	DNMT3L, ICOSLG
rs734252	none	MAN2A2	MAN2A2, FURIN, FES
rs738322	PLA2G6 (intron)	PLA2G6	PLA2G6, TMEM, MAFF
rs7578035	none	none	MGAT4A, C2orf55
rs7585265	RPE (3′ UTR)	RPE, C2orf67	RPE, C2orf67, UNC80
rs7734434	none	none	none
rs7965349	OASL (intron)	OASL	OASL, C12orf43, HNF1A
rs8014856	GALC (intron)	GALC	GALC, GPR65
rs8015102	GALC (intron)	GALC	GALC, GPR65
rs8027781	none	none	DENND4A
rs8192917	GZMB (exon)	GZMB	GZMB, GZMH
rs851984	ESR1 (intron)	ESR1	ESR1
rs859648	none	none	none
rs881375	none	none	TRAF, PHF19
rs907613	LSP1 (5′ UTR)	LSP1	LSP1, SYT8, TNNI2, mir-4298
rs9271170	none	none	HLA-DRB4, HLA-DQA1, HLA-DQB1
rs9271612	none	none	HLA-DRB4, HLA-DQA1, HLA-DQB1
rs9271613	none	none	HLA-DRB4, HLA-DQA1, HLA-DQB1
rs9303029	none	C17orf62	C17orf62, NARF, C17orf101, HEXDC, LOC100505970
rs9314099	FAM172A (5′ UTR)	FAM172A	FAM172A, KIAA0825
rs9314100	FAM172A (5′ UTR)	FAM172A	FAM172A, KIAA0825
rs9393984	none	none	HCG8, HCG9, HLA-A, HLA-A29.1
rs9442234	FAM131C (intron)	FAM131C	FAM131C, CLCNKB, CLCNKA, HSPB7
rs9909593	IKZF3 (intron)	IKZF3	IKZF3

7.8 Genes Located Near Hit-SNPs for GATA3 in Th2 cells

Table 7.6: Genes located in or near SNP-hits for GATA3 in Th2 Cells. We recorded any genes containing a hit-SNP for GATA3 in addition to any genes within 2kbp of a hit-SNP and any genes within 50kbp of a hit-SNP.

rsID of Hit-SNP	Genes overlapping Hit-SNP	Genes with 2kbp of Hit-SNP	Genes within 50kbp of Hit-SNP
rs1006353	none	none	MTIF3, GTF3A
rs10152591	none	none	none
rs10267797	none	none	DLD, LAMB1
rs10516286	C1QTNF7 (5′)	C1QTNF7	C1QTNF7
rs10899476	GAB2 (intron)	GAB2	GAB2
rs10929322	none	none	ATG16L1, INPP5D
rs10930310	STK39 (intron)	STK39	STK39
rs10995195	ZNF365 (intron)	ZNF365	ZNF365
rs11130317	GLT8D1 (intron)	GLT8D1	GLT8D1, GNL3, PBRM1, SPCS1, NEK4
rs11135484	ERAP2 (intron)	ERAP2	ERAP2, LNPEP
rs11142	SORT1 (exon)	SORT1	SORT1, PSMA5, MVBPHL
rs11663558	NPC1 (intron)	NPC1	NPC1, ANKRD29, C18orf8
rs11692725	none	none	none
rs12132349	C1orf106 (intron)	C1orf106	C1orf106, GPR25, CAMSAP1L1
rs12252379	none	none	LOC254312
rs1279750	none	none	TRIM7, OR2V2
rs1285403	none	none	none
rs12942708	none	none	PEMT
rs12991737	none	none	IL18RAP, IL18R1, mir-4772
rs1323292	none	none	RGS1
rs1327235	none	none	none
rs1388387	THADA (intron)	THADA	THADA
rs1406069	KIAA1267 (intron)	KIAA1267	KIAA1267
rs1420106	none	IL18RAP	IL18RAP, IL18R1, mir-4772
rs1596017	none	none	ATG16L1, INPP5D, SCARNA5, SCARN6
rs1610588	none	none	none
rs17103930	none	none	none
rs17753121	none	none	GTF3A, MTIF3
rs1859156	BMPR1B	BMPR1B	BMPR1B
rs1886730	TNFRSF14 (intron)	TNFRSF14	TNFRSF14, MMEL1, C1orf93, HES5, PANK4

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs1893447	GAB2 (intron)	GAB2	GAB2, USP35
rs1983351	SLCO3A1 (intron)	SLCO3A1	SLCO3A1
rs2045592	AFF1 (intron)	AFF1	AFF1
rs2058622	IL18R1 (intron)	IL18R1	IL18R1, IL1RL1, IL18RAP
rs2106346	TSC1 (intron)	TSC1	TSC1, C9orf9, AK8
rs2145998	none	none	ZMIZ1, PPIF, ZCCHC24
rs2239826	none	none	CCR6, FGFR1
rs2275271	CNNM2 (exon)	CNNM2	CNNM2, NT5C2
rs2278563	ARID5A (intron)	ARID5A	ARID5A, KIAA1310
rs2287900	none	CAPSL	CAPSL, UGT3A1, IL7R
rs231727	none	none	CTLA4
rs2408025	SLC38A9 (intron)	SLC38A9	SLC38A9
rs281438	none	ICAM4, ICAM5	ICAM4, ICAM5, ICAM1, MRPL4, ZGLP1, FDX1L, RAVER1, ICAM3
rs286913	EHF (intron)	EHF	EHF
rs2984920	none	RGS1	RGS1
rs3129763	none	none	HLA-DQB1, HLA-DQA1, HLA-DRB4
rs34251816	WHSC1 (intron)	WHSC1	WHSC1, LETM1
rs34361	POC5 (intron)	POC5	POC5, ANKDD1B
rs3732123	none	none	IL1RL1, IL18R1, IL18RAP, mir-4772
rs3758253	HEMGN (5′)	HEMGN	HEMGN, ANP32B, C9orf156
rs3758354	none	none	ANXA1
rs3792112	ATG16L1 (intron)	ATG16L1	ATG16L1, SAG, SCARNA6, SCARNA5
rs4024110	CAPSL (intron)	CAPSL	CAPSL, UGT3A1, IL7R
rs40452	RIPK2 (intron)	RIPK2	RIPK2
rs41156	MTMR3 (intron)	MTMR3	MTMR3
rs4522587	ICOS (intron)	ICOS	ICOS
rs460901	none	none	SSBP2
rs4668070	CERS6 (intron)	CERS6	CERS6, mir-4774
rs4679048	none	none	XYLB
rs4683205	none	none	CCR3
rs4772190	UBAC2 (intron)	UBAC2	UBAC2, FKSG29, mir-623
rs4939490	none	none	CD6
rs507101	SLC3A2 (intron)	SLC3A2	SLC3A2, WDR74, RNU2-2, SNHG1, SNORD22, 26, 28, 30, 25, 27, 31 and 25
rs6441957	none	none	CCR3
rs6478486	none	none	TRAF1, PHF19, PSMD5
rs6556405	RNF145 (5′ UTR)	RNF145	RNF145

Hit-SNP	Overlapping	Within 2kbp	Within 50kbp
rs6604026	RPL5 (intron)	RPL5, SNORD21	RPL5, SNORD21, FAM69A, EVI5, SNORD66
rs6685472	none	none	none
rs6936240	PDE7B (intron)	PDE7B	PDE7B
rs7004723	ASAP1 (intron)	ASAP1	ASAP1
rs7165988	C15orf53 (exon)	C15orf53	C15orf53
rs7171233	none	C15orf53	C15orf53
rs727263	UBAC2 (intron), FKSG29	UBAC2, FKSG29	UBAC2, FKSG29, GPR183, mir-623
rs7488228	none	none	none
rs7578035	none	none	MGAT4A, C2orf55
rs7585265	RPE (3′ UTR)	RPE, C2orf67	RPE, C2orf67, UNC80
rs7602799	none	none	none
rs7705826	none	SLC22A5	SLC22A5, C5orf56, mir-3936
rs7988412	GTF3A (intron)	GTF3A	GTF3A, MTIF3
rs8027781	none	none	DENND4A
rs8062727	none	none	NOD2, SNX20, NKD1
rs809930	LOC100287063 (intron)	LOC100287063	LOC100287063
rs851984	ESR1 (intron)	ESR1	ESR1
rs881375	none	none	TRAF, PHF19
rs907613	LSP1 (5′ UTR)	LSP1	LSP1, SYT8, TNNI2, mir-4298
rs9271612	none	none	HLA-DRB4, HLA-DQA1, HLA-DQB1
rs9271613	none	none	HLA-DRB4, HLA-DQA1, HLA-DQB1
rs933672	SLC38A9 (intron)	SLC38A9	SLC38A9
rs9869432	LRRFIP2 (intron)	LRRFIP2	LRRFIP2

7.9 Summary of Genomic Annotation

Table 7.7: Summary of genomic annotation. Summary of information for hit-SNPs found for T-bet, GATA3 in Th1 cells and GATA3 in Th2 cells is shown. For Modifications column, Promoter = H3K4me3 mark present, Enhancer = H3K4me1 mark present, Both = H3K4me3 and H3K4me1 marks both present. For DNase column, 'All' indicates a region of DNase hypersensitivity was found to overlap all transcription factor events at that hit-SNP. In some cases, DNase hypersensitivity was only found in Th1 or Th2 cells whereas binding was seen in both Th1 and Th2 cells. These cases are indicated by 'Th1 only' and 'Th2 only' respectively. For Motifs column, A = Motif Altered, P= Motif Present, D = Motif found within 20bp either side of hit-SNP by EMBOSS `dreg` tool, F= Motif found within 100bp either side of hit-SNP by FIMO. For final column, nsSNP = nonsynonymous SNP. In 'Associated Traits' column abbreviations used are: UC = Ulcerative Colitis, RA = Rheumatoid Arthritis, MS = Multiple Sclerosis, AS = Ankylosing Spondylitis, SLE = Systemic Lupus Erythematosus, PBC = Primary Biliary Cirrhosis, PSP = Progressive Supranuclear Palsy, ADHD = Attention Deficit Hyperactivity Disorder, COPD - Chronic Obstructive Pulmonary Disease, SS = Systemic Sclerosis, CLL = Chronic Lymphocytic Leukaemia, T1D = Type 1 Diabetes, T2D = Type 2 Diabetes.

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs1006353	T-bet, Th1Gata3, Th2Gata3	Body mass index	none	none	none	T-bet(A.D), GATA3(P.D&A.F)	no
rs10152590	T-bet, Th1Gata3	Height	none	Promoter	all	T-bet(A.D)	no
rs10152591	T-bet, Th1Gata3, Th2Gata3	Height	none	Promoter	Th1 only	none	no
rs10263639	Th1Gata3	Breast cancer	none	none	no	GATA3(P.D)	no
rs10267797	Th1Gata3, Th2Gata3	UC	none	none	all	none	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs10421601	Th1Gata3	Diabetic retinopathy	none	Enhancer	all	GATA3(P.F)	no
rs10516286	Th2Gata3	Conduct disorder	C1QTNF7	none	all	none	Yes
rs10760294	T-bet	Polycystic ovary syndrome	DENND1A	none	none	none	no
rs10808568	T-bet	Coeliac	none	none	all	T-bet(P.F)	no
rs10828247	Th1Gata3	Meningioma	MLLT10	Promoter	all	none	no
rs10850435	T-bet	HDL cholesterol	MVK	Enhancer	all	none	Yes
rs10861892	T-bet, Th1Gata3	Conduct disorder	CMKLR1	none	none	none	no
rs10899476	Th2Gata3	Menarche, Alzheimer's	GAB2	none	none	GATA3(P.D)	no
rs10929322	T-bet, Th1Gata3, Th2Gata3	Crohn's	none	none	all	GATA3(P.D)	Yes
rs10930310	T-bet, Th1Gata3, Th2Gata3	Blood pressure	STK39	Both	all	none	no
rs10995195	T-bet, Th1Gata3, Th2Gata3	Breast cancer	ZNF365	Enhancer	all	none	no
rs11009175	T-bet	Depression	none	Promoter	all	none	no
rs11082995	T-bet	Protein quantitative trait loci	OSBPL1A	none	none	none	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs11130317	T-bet, Th1Gata3, Th2Gata3	Bipolar and mood disorders	GLT8D1	none	none	GATA3(P.D)	Yes
rs11135484	T-bet, Th1Gata3, Th2Gata3	Crohn's	ERAP2	Both	all	T-bet(P.D & P.F), GATA3(P.D)	no
rs11142	Th1Gata3, Th2 Gata3	Chronic kidney disease	SORT1	none	none	none	no
rs11571293	T-bet	T1D, RA	none	Promoter	all	none	no
rs11663558	Th2Gata3	Obesity	NPC1	none	none	GATA3(P.D)	Yes
rs11692725	Th2Gata3	MS, Cardiographic traits	none	none	none	none	no
rs11742570	Th1Gata3	Crohn's, UC, MS	none	Promoter	all	none	no
rs12038333	Th1Gata3	Macular Degeneration	CFH	none	none	none	no
rs12132349	Th1Gata3, Th2Gata3	UC, Crohn's, AS, Coeliac	C1orf106	Enhancer	all	none	no
rs12136874	T-bet	Knee osteoarthritis	none	none	none	none	no
rs12252379	Th2Gata3	Uterine Fibroids	none	none	none	none	no
rs12574073	T-bet	SLE	none	none	none	none	no
rs12653750	T-bet	Asthma	RAD50	none	all	none	no
rs1279750	Th1Gata3, Th2Gata3	Platelet counts	none	none	all	GATA3(A.D&P.F)	no
rs1285403	Th2Gata3	Protein quantitative trait loci	none	none	none	GATA3 (A.D)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs12928552	T-bet	Coronary heart disease	NLRC5	Promoter	all	none	no
rs12942708	Th2Gata3	Coronary heart disease	none	none	none	none	no
rs12946510	T-bet	T1D, UC, RA, Crohn's, Asthma, PBC	none	Enhancer	none	none	Yes
rs12991737	Th1Gata3, Th2Gata3	Asthma	none	none	Th2 only	GATA3(A.D)	no
rs1323292	Th1Gata3, Th2Gata3	Coeliac	none	none	all	T-bet(P.F), GATA3(P.F)	no
rs1327235	Th2Gata3	Blood pressure	none	none	none	none	no
rs133015	Th1Gata3	Cutaneous nevi	PLA2G6	Enhancer	none	none	no
rs133016	Th1Gata3	Cutaneous nevi	PLA2G6	Enhancer	none	none	no
rs133017	Th1Gata3	Cutaneous nevi	PLA2G6	Enhancer	none	T-bet(P.F)	no
rs13333528	T-bet	Colorectal cancer	CDH1	none	all	T-bet(P.D)	no
rs1335512	T-bet, Th1Gata3	Cutaneous nevi	none	none	all	none	no
rs13418039	Th1Gata3	Sudden cardiac arrest	none	Both	none	none	no
rs1364340	T-bet, Th1Gata3	Obesity	none	none	none	none	no
rs1374910	T-bet	T2D	IGF2BP2	none	none	none	no
rs1388387	Th2Gata3	T2D	THADA	none	all	GATA3(A.D)	Yes
rs1406069	Th1Gata3, Th2Gata3	PSP, Parkinson's	KIAA1267	none	none	GATA3(P.F)	Yes

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs1420106	T-bet, Th1Gata3, Th2Gata3	Coeliac, Crohn's	IL18RAP	Promoter	all	GATA3(P.D&P.F)	no
rs1445899	Th1Gata3	T1D	CAPSL	none	none	none	Yes
rs1465321	T-bet, Th1Gata3	Coeliac, Crohn's	IL18R1	none	none	T-bet (P.D)	no
rs1481892	Th1Gata3	Menarche	ARNTL	none	none	T-bet(P.F)	no
rs149299	T-bet	T1D, Crohn's	none	none	none	T-bet(P.F)	no
rs1551398	T-bet, Th1Gata3	Crohn's	none	Enhancer	none	none	no
rs1551399	T-bet, Th1Gata	Crohn's	none	Enhancer	none	none	no
rs1570069	Th1Gata3	Phospholipid levels	ELOVL2	none	none	T-bet(P.F)	no
rs1596017	Th1Gata3, Th2Gata3	Crohn's	none	Enhancer	none	GATA3(P.D)	Yes
rs1610588	T-bet, Th1Gata3, Th2Gata3	Nasopharyngeal carcinoma	none	Promoter	all	none	no
rs1631457	T-bet	Tonometry	SYNE1	none	none	none	no
rs17032405	T-bet	Height	none	none	none	none	no
rs17035375	Th1Gata3	Coeliac	PLEK	none	none	none	Yes
rs17035378	Th1Gata3	Coeliac	PLEK	none	none	none	Yes

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs17103930	T-bet, Th1Gata3, Th2Gata3	Conduct disorder	none	Both	Th1 only	T-bet(P.D)	no
rs17234657	T-bet	Crohn's	none	none	none	none	no
rs1738074	T-bet	MS, Coeliac	TAGAP	Promoter	all	none	no
rs17411033	Th1Gata3	UC	none	none	none	GATA3(P.D)	no
rs17533167	Th1Gata3	Follicular lymphoma	none	Enhancer	all	GATA3 (A.D)	no
rs17753121	Th2Gata3	Weight	none	none	none	GATA3(P.D)	no
rs17772411	T-bet	Colorectal cancer	CDH1	Enhancer	all	none	no
rs1788183	T-bet	Atrioventricular conduction	none	none	none	T-bet(P.F)	Yes
rs1806689	T-bet	Hypertension	none	none	all	none	no
rs1859156	Th1Gata3, Th2Gata3	ADHD	BMPR1B	Enhancer	none	none	no
rs1886730	Th1Gata3, Th2Gata3	UC	TNFRSF14	Both	Th1 only	GATA3 (A.D & P.F)	no
rs1893447	Th2Gata3	Menarche, Alzheimer's	GAB2	none	none	none	no
rs1918788	T-bet	Parkinson's, PSP	KIAA1267	Promoter	none	T-bet(P.D & P.F)	Yes
rs1983351	Th2Gata3	Inattentive symptoms	SLCO3A1	none	none	none	no
rs2016755	Th1Gata3	Cutaneous nevi	PLA2G6	Enhancer	none	GATA3(P.D)	no
rs2045592	Th2Gata3	Triglycerides	AFF1	none	none	GATA3(P.F)	no
rs2058622	Th1Gata3, Th2Gata3	Coeliac, Crohn's	IL18R1	Both	Th2 only	none	no
rs2070615	T-bet	Bipolar	CACNB3	none	none	T-bet(P.D & P.F)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs2106346	T-bet, Th1Gata3, Th2Gata3	Psoriasis	TSC1	Enhancer	none	none	no
rs2145998	Th1Gata3, Th2Gata3	Height	none	none	none	none	no
rs216172	T-bet, Th1Gata3	Coronary heart disease	SMG6	Enhancer	all	T-bet(P.D & P.F)	no
rs2179225	T-bet	Knee osteoarthritis	none	none	all	T-bet(P.D)	no
rs2239826	Th2Gata3	RA	none	none	all	none	no
rs2267020	Th1Gata3	Height	BCR	Enhancer	none	GATA3(P.D)	no
rs2268080	T-bet	Frekling, Skin sensitivity, Hair colour	RALV	none	all	T-bet(P.D)	no
rs2273017	T-bet	Graves' disease	C6orf10	none	none	T-bet(P.D)	no
rs2275271	Th1Gata3, Th2Gata3	Schizophrenia	CNNM2	Both	Th1 only	GATA3(P.D)	no
rs2278563	Th1Gata3, Th2Gata3	Brain oscillations	ARID5A	Enhancer	none	none	no
rs2282030	Th1Gata3	Mortality in heart failure	PMSC1	none	none	T-bet(P.F)	Yes
rs228599	T-bet	Response to metformin	ATM	Promoter	all	T-bet(A.D & A.F)	Yes
rs2287900	Th2Gata3	UC	CAPSL	none	none	GATA3(P.F)	Yes
rs2288344	T-bet	Keloid	NEDD4	Enhancer	all	none	no
rs231727	Th1Gata3, Th2Gata3	Graves' disease, Alopecia areata	none	none	none	GATA3(P.D)	Yes

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs2331903	T-bet	PSP	STX6	none	all	none	no
rs2387397	T-bet	RA	none	Promoter	all	T-bet(A.D)	no
rs2408025	Th2Gata3	Height	SLC38A9	none	none	T-bet(P.F), GATA3(P.D)	Yes
rs2421016	T-bet	Height	PLEKHA1	Promoter	none	T-bet(P.D)	no
rs2532234	T-bet, Th1Gata3	PSP, Parkinson's Coelic	KIAA1267	Promoter	all	none	Yes
rs2532296	Th1Gata3	PSP, Parkinson's	KIAA1267	Enhancer	none	T-bet(P.F)	Yes
rs261360	Th1Gata3	Hair morphology	none	Both	all	GATA3(P.F)	no
rs2639186	Th1Gata3	Processing speed	none	none	none	none	no
rs264834	T-bet	Protein quantitative trait loci	DOCK2	none	none	none	no
rs2702136	Th1Gata3	Working memory	ITSN2	Promoter	none	GATA3(P.F)	no
rs2703078	T-bet, Th1Gata3	Coeliac	none	Enhancer	all	T-bet(A.D)	no
rs2706383	T-bet, Th1Gata3	Fibrinogen	C5orf56	Both	all	none	no
rs2720665	T-bet	Hodgkin's lymphoma	PVT1 oncogene	Enhancer	none	none	no
rs2762060	T-bet, Th1Gata3	MS	none	Enhancer	all	T-bet(P.D)	no
rs281438	Th1Gata3, Th2Gata3	Soluble ICAM-1	ICAM4, ICAM5	Promoter	all	GATA3(A.D)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs28365932	T-bet	Conduct disorder	DDX54, C12orf52	Promoter	all	T-bet(P.D)	no
rs2865530	Th1Gata3	Pulmonary function	CFDP1	Both	all	GATA3(P.D)	no
rs286913	Th1Gata3, Th2Gata3	Antipsychotic response	EHF	none	none	none	no
rs2897908	T-bet	Hypertension	none	none	all	none	no
rs2904259	T-bet	COPD	FAM13A	none	all	T-bet(P.D)	no
rs2984920	T-bet, Th2Gata3	Coeliac	RGS1	Promoter	all	T-bet(P.D)	no
rs3024505	T-bet	T1D, UC, Crohn's	IL10	none	all	none	no
rs307896	T-bet	MS	SAE1	Both	all	T-bet(P.F)	no
rs3091310	T-bet	Coeliac	CCR3	Promoter	all	none	no
rs3101018	Th1Gata3	Lung cancer, SLE	CLIC1, MSH5	Enhancer	all	none	Yes
rs3129763	Th1Gata3, Th2Gata3	SS, Leprosy, CLL	none	Enhancer	all	none	Yes
rs3131096	T-bet, Th1Gata3	Lung adenocarcinoma	TPK1	Promoter	all	T-bet(A.D)	Yes
rs3181080	T-bet	Coeliac	CCR1	Enhancer	all	none	no
rs3213621	Th1Gata3	Keloid	ZNYMD10, RASSF1	Both	none	none	Yes
rs34251816	Th2Gata3	Coronary heart disease	WHSC1	none	none	GATA3(P.D)	Yes
rs34361	Th2Gata3	Body mass index	POC5	none	none	GATA3(P.D&P.F)	no
rs354042	T-bet	MS	none	Enhancer	none	T-bet(P.F)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs36594	T-bet	Lung cancer	MTMR3	none	none	none	no
rs3732123	Th1Gata3, Th2Gata3	White blood cell types	none	none	Th2 only	none	no
rs3733041	T-bet	Major mood disorders	GLT8D1	none	none	none	Yes
rs3738398	Th1Gata3	Neurocognitive phenotypes	DISC1	Promoter	all	none	no
rs3752948	T-bet	Endothelial function traits	OBFC1	none	all	none	no
rs3758253	Th1Gata3, Th2Gata3	Quantitative traits	HEMGN	Enhancer	Th1 only	T-bet(P.F)	Yes
rs3758354	T-bet, Th1Gata3, Th2Gata3	Schizophrenia, Bipolar, Depression	none	none	all	none	no
rs3784388	T-bet	Common traits (other)	AP3S2	Promoter	all	none	no
rs3788013	Th1Gata3	T1D	UBASH3A	Enhancer	all	T-bet(P.F)	no
rs3789053	T-bet	Cognitive performance	MDM4	Promoter	none	T-bet(P.F)	Yes
rs3792112	Th2Gata3	Crohn's	ATG16L1	none	all	none	Yes
rs3806288	Th1Gata3	Diabetic retinopathy	MRPS15, CSF3R	Both	all	GATA3(A.D)	no
rs3810610	T-bet	MS	MAPK1	none	none	none	no
rs3821383	T-bet	MS, Coeliac, Colon cancer, Telomere length	MYNN	Promoter	none	none	Yes
rs3863439	Th1Gata3	Breast cancer	none	none	none	GATA3(P.D)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs4024110	T-bet, Th2Gata3	UC	CAPSL	none	all	none	Yes
rs40452	Th1Gata3, Th2Gata3	Leprosy	RIPK2	none	all	none	no
rs41156	Th2Gata3	T1D	MTMR3	none	none	T-bet(P.F)	no
rs4265380	T-bet	Coeliac, AS	RUNX3	Enhancer	all	none	Yes
rs4398410	T-bet	Lung adenocarcinoma	none	none	none	none	no
rs4432068	T-bet, Th1Gata3	SLE	none	none	all	none	no
rs4522587	Th1Gata3, Th2Gata3	Coeliac	ICOS	Enhancer	all	GATA3(P.F)	no
rs4593512	T-bet	Antipsychotic response	none	none	none	none	no
rs460901	Th1Gata3, Th2Gata3	Anorexia nervosa	none	Promoter	all	GATA3(P.D)	no
rs4667114	Th1Gata3	Coeliac, AS	none	none	none	none	no
rs4668070	T-bet, Th1Gata3, Th2Gata3	TNF response	CERS6	none	none	GATA3(A.D)	no
rs4679048	Th1Gata3, Th2Gata3	Cardiographic conduction	none	Enhancer	Th1 only	GATA3(P.F)	no
rs4683205	T-bet, Th1Gata3, Th2Gata3	Brain structure	none	none	none	GATA3(P.D)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs4686760	Th1Gata3	Plasma vWF and FVIII levels	none	none	none	T-bet(P.F), GATA3(A.D&P.F)	no
rs4713186	T-bet	Height, Platelet aggregation	C6orf100	Promoter	all	none	Yes
rs4713581	Th1Gata3	Graves' disease, SS, RA	none	none	none	none	no
rs4713582	Th1Gata3	Graves' disease, SS, RA	none	none	none	none	no
rs4772190	Th1Gata3, Th2Gata3	Basal cell carcinoma	UBAC2	Enhancer	Th2 only	T-bet(P.F)	no
rs4847378	T-bet	Bipolar, Schizophrenia	FAM69A	Promoter	all	none	no
rs485499	T-bet	PBC	none	none	all	none	no
rs485789	T-bet	PBC	none	none	all	T-bet(A.F)	no
rs4888376	Th1Gata3	Pulmonary function	none	Enhancer	all	none	no
rs4906172	T-bet	Menopause	DYNC1H1	none	all	T-bet(P.F)	no
rs4917014	Th1Gata3	SLE	none	Promoter	all	none	no
rs4939490	Th2Gata3	MS	none	Enhancer	all	none	no
rs4944195	T-bet	Menarche, Alzheimer's	GAB2	none	all	none	no
rs507101	Th1Gata3, Th2Gata3	Antipsychotic response	SLC3A2	none	Th1 only	T-bet(P.F), GATA3(P.D)	no
rs5778	T-bet	Body Mass	TRHR	none	none	T-bet(A.D)	no
rs631106	T-bet	Cholesterol, Triglycerides	USP1	Promoter	all	none	no
rs6441957	Th1Gata3, Th2Gata3	Brain structure	none	none	none	GATA3(P.D & P.F)	no
rs6478486	Th2Gata3	Coeliac, RA	none	Enhancer	all	GATA3(P.D&P.F)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs6556405	T-bet, Th2Gata3	Mean platelet volume	RNF145	Promoter	all	none	no
rs6604026	Th2Gata3	MS	RPL5, SNORD21	none	none	none	no
rs6651252	Th1Gata3	Ovarian cancer, Crohn's	none	Promoter	all	none	no
rs6685472	Th2Gata3	Heart failure	none	none	none	GATA3(P.D&P.F)	Yes
rs6686643	Th1Gata3	Total ventricular volume	MGST3	Enhancer	none	GATA3(P.D&P.F)	no
rs6723449	T-bet	UC	CXCR2	none	none	none	no
rs6784841	T-bet	Coeliac	FRMD4B	none	all	T-bet(P.D)	no
rs6919110	Th1Gata3	Height	FAM8A1	Promoter	all	GATA3(P.D)	Yes
rs6936240	Th2Gata3	F-cell distribution	PDE7B	Promoter	all	none	Yes
rs6997	Th1Gata3	Primary sclerosing cholangitis, Crohn's, UC	TCTA, AMT	Enhancer	none	none	Yes
rs7004723	Th1Gata3, Th2Gata3	MS	ASAP1	Promoter	none	GATA3(P.D)	no
rs707952	Th1Gata3	SS, Leprosy, CLL, RA, Schizophrenia	HLA-DQA1	none	none	2*T-bet (P.F&P.F)	Yes
rs7090445	Th1Gata3	Acute lymphoblastic leukaemia	ARID5B	Both	all	GATA3(A.D)	no
rs7117353	Th1Gata3	Suicide attempts in bipolar disorder	C11orf41	none	none	none	no
rs7156191	T-bet	Schizophrenia	TRAF3	none	none	none	no
rs7165988	Th2Gata3	Psychological traits	C15orf53	Both	all	GATA3(P.D)	Yes
rs7171233	Th2Gata3	Psychological traits	C15orf53	Both	all	none	Yes

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs727263	T-bet, Th2Gata3	Basal cell carcinoma	UBAC2, FKSG29	none	all	T-bet(P.D), GATA3(P.D)	no
rs7283760	Th1Gata3	Coeliac	none	none	none	none	no
rs734252	Th1Gata3	ADHD	MAN2A2	Promoter	all	none	no
rs738322	Th1Gata3	Cutaneous nevi	PLA2G6	none	none	GATA3(P.F)	no
rs743776	T-bet	T1D, RA	none	none	none	T-bet(A.D &P.F)	no
rs743777	T-bet	T1D, RA	none	none	none	T-bet(P.F)	no
rs7441808	T-bet	T1D, RA	none	none	all	T-bet(A.D)	no
rs7488228	Th2Gata3	Optic disc size (cup)	none	none	all	GATA3(P.D)	no
rs7578035	Th1Gata3, Th2Gata3	Bipolar disorder	none	none	Th1 only	none	no
rs7585265	Th1Gata3, Th2Gata3	Serum metabolites	RPE, C2orf67	none	none	2*T-bet (P.F&P.F), Gata3(P.D)	Yes
rs7602799	Th2Gata3	Atrioventricular conduction	none	Promoter	none	none	no
rs7705826	Th2Gata3	Crohn's	SLC22A5	Enhancer	all	none	no
rs7734434	Th1Gata3	Crohn's	none	Promoter	all	none	no
rs7787612	T-bet	Alcohol consumption	AUTS2	none	none	none	no
rs7801282	T-bet	Aortic root size	CRCP	Promoter	none	none	Yes
rs7850707	T-bet	Height	none	none	all	T-bet(P.F)	Yes
rs7859727	T-bet	Heart conditions	CDKN2B antisense RNA	none	all	GATA3(P.F)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs7923800	T-bet	Common traits (other)	KIAA1267	none	none	none	no
rs7965349	Th1Gata3	T2D	OASL	Enhancer	all	none	no
rs7988412	Th2Gata3	Body mass index	GTF3A	Promoter	all	none	no
rs8008961	T-bet	PBC	RAD51B	Promoter	none	none	no
rs8014856	Th1Gata3	MS	GALC	none	none	none	no
rs8015102	Th1Gata3	MS	GALC	none	none	None	no
rs8027781	Th1Gata3, Th2Gata3	Haematological traits	none	none	all	GATA3(P.D&P.F)	no
rs8062727	T-bet, Th2Gata3	Leprosy	none	Enhancer	all	T-bet(A.D)	no
rs809930	Th2Gata3	Blood pressure	LOC100287063	Enhancer	all	T-bet(P.F), GATA3(P.F)	Yes
rs8129743	T-bet	Oesophageal cancer	Runx1	none	all	none	no
rs8192917	Th1Gata3	Vitiligo	GZMB	Promoter	none	GATA3(P.D)	Yes
rs851984	Th1Gata3, Th2Gata3	Bone mineral density	ESR1	none	Th2 only	T-bet(P.F), GATA3(P.D)	no
rs859648	Th1Gata3	Coeliac	none	none	none	GATA3(P.D&P.F)	no
rs881375	Th1Gata3, Th2Gata3	Coeliac, RA	none	Promoter	all	GATA3(P.D&P.F)	no
rs907613	T-bet, Th1Gata3, Th2Gata3	UC	LSP1	Both	all	none	no
rs9271170	Th1Gata3	SLE	none	Both	all	T-bet (P.F), GATA3(P.D)	no

Hit-SNP rsID	Binding Site for?	Associated Traits	Gene(s) within 2kbp	Modifications	DNase	Motifs	LD with nsSNP?
rs9271612	Th1Gata3, Th2Gata3	SS, Leprosy, CLL	none	none	Th2 only	none	no
rs9271613	Th1Gata3, Th2Gata3	SS, Leprosy, CLL	none	none	Th2 only	none	no
rs9292777	T-bet	Crohn's, UC, MS	none	Enhancer	all	none	no
rs9300536	T-bet	Basal cell carcinoma	UBAC2	none	all	T-bet(P.F)	no
rs9303029	T-bet, Th1Gata3	Protein quantitative trait loci	C17orf62	Promoter	all	none	no
rs9314099	Th1Gata3	Diabetic retinopathy	FAM172A	Promoter	all	GATA3(P.D)	no
rs9314100	Th1Gata3	Diabetic retinopathy	FAM172A	Promoter	all	GATA3(P.D)	no
rs933672	Th2Gata3	Height	SLC38A9	none	none	T-bet(P.F)	Yes
rs9393984	T-bet, Th1Gata3	Liver injury, Vitiligo	none	Promoter	all	none	Yes
rs9442234	Th1Gata3	Cognitive performance	FAM131C	none	none	none	no
rs9594738	T-bet	Bone mineral density	none	Enhancer	all	none	no
rs9869432	Th2Gata3	P-tau181p	LRRFIP2	Enhancer	all	none	no
rs9909593	T-bet, Th1Gata3	T1Ds, UC, Crohn's, RA, Asthma, PBC	IKZF3	Promoter	all	T-bet(A.D & P.F)	Yes
rs9944207	T-bet	Height	SNUPN	Enhancer	none	none	no